2020-06

Data driven approach for predicting student dropout in secondary schools

Mduma, Neema

NM-AIST

DATA DRIVEN APPROACH FOR PREDICTING STUDENT DROPOUT IN SECONDARY SCHOOLS

Neema Mduma

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

June, 2020

ABSTRACT

Student dropout is among the challenges that face most schools in developing countries particularly in Africa. In Tanzania alone, student dropout in secondary schools is pronounced to be around 36%. In addressing the student dropout problem, a thorough understanding of the fundamental factors that cause the student dropout is essential. Several researchers have identified and proposed causes, methods and strategies that will help to reduce or stop the student dropout problem, however, most of the proposed solutions didn't show promising results and the students dropout trend continue to increase over time. This study focused on developing a data driven approach that will help to identify and predict students who are at risk of dropping out of school in order to facilitate an intervention program as an active measure in eliminating the problem of dropout in Tanzania. In doing so, (a) 122 research articles were examined, (b) 4 focus group discussions and 2 round table surveys with 38 respondents from 5 districts (Arusha, Mbeya, Kisarawe, Rufiji and Nzega) were conducted, and (c) 3 datasets from Tanzania and India were used in order to identify factors that contribute significantly to student dropout problem, disclose the best classifier from the commonly used classifiers (Logistic Regression, Random Forest, K-nearest Neighbor and Multilayer Perceptron) and assessing the data balancing techniques for predictive performance of the model. Results revealed that, most of the respondents mentioned students' gender, age, parent's income, number of qualified teachers and remoteness as the main contributing factors to the students' dropout problem in secondary schools. Furthermore, results from the examined articles indicated that, most studies conducted in developing countries focused on the social aspects of student dropout, and a paltry mentioned the use of other approaches such as machine learning. Nevertheless, results from data driven approach development shows that the Logistic Regression and Multilayer perceptron achieved the highest performance when over-sampling technique was employed. Also, the hyper parameter tuning improved the algorithm's performance compared to its baseline settings, and stacking of the classifiers improved the overall predictive performance of the developed approach. The study, therefore, recommends the developed approach to be considered by relevant authorities in identifying and predicting students at risk of dropping out for early intervention, planning and informative decisions making on addressing the student dropout problem.

DECLARATION

I, Neema Mduma do hereby declare to the Senate of Nelson Mandela African Institution of Science and Technology that this dissertation is my own original work and that it has neither been submitted nor presented for similar degree award in any other institution.

Neema Mduma	Date
(Candidate)	

The above declaration is confirmed

Dr. Khamisi Kalegele	Date
(Supervisor 1)	
Dr. Dina Machuve	Date
(Supervisor 2)	

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and found the dissertation titled "Data Driven Approach for Predicting Student Dropout in Secondary Schools" conforming to the standard and format acceptable by the Nelson Mandela African Institution of Science and Technology.

Dr. Khamisi Kalegele
(Supervisor 1)

Date

Dr. Dina Machuve

(Supervisor 2)

Date

ACKNOWLEDGMENTS

Many people have walked with me in this adventure and they surely deserve my greatest gratitude. First, I would like to thank the almighty God for being my present help in times of need and a strong tower to turn to, He is the Rock of my salvation. I thank the African Development Bank (AfDB) for the generous scholarship to pursue my studies at the Nelson Mandela African Institution of Science and Technology (NM-AIST). My sincere thanks to my supervisors: Dr. Khamisi Kalegele of the Tanzania Commission for Science and Technology (COSTECH), and Dr. Dina Machuve of NM-AIST. Thank you for your excellent guidance, patience, motivation, encouragement and for providing an excellent atmosphere for doing this study.

I offer my sincerest gratitude to my mentor the founding Vice Chancellor of the Nelson Mandela African Institution of Science and Technology, Prof. Burton Mwamila, for encouraging me during my studies and allowing me to grow as a research scientist. I extend my special thanks to the management of NM-AIST for giving me permission to pursue my studies and for their continue support during the whole period of my studies. Beside NM-AIST management, I thank the rest of the staff members of the school of Computational and Communication Science and Engineering (CoCSE) at NM-AIST for their insightful comments and questions that awakened my brain to think.

I would like to appreciate the President Office Regional Administration and Local Government (PORALG), Data for Local Impact (DLi) and Eagle Analytics for their support. Without them, I wouldn't have succeeded to collect data for conducting this study. My sincere thanks to my parents, Rev. and Mrs. Mathias Mdumah, together with Eng. and Mrs. Simon Sumari, for always being there to counsel and for their honest prayers. Thanks to my fellow colleagues just to mention a few, Hudson Laizer, who made this journey possible in one way or another. Lastly, may the Lord reward all those people whose names I have not mentioned here for the support during this journey. You played a significant role and I am grateful to you all.

DEDICATION

I dedicate this dissertation to my father and greatest friend, Rev. Mathias Mdumah and my dearest grandfather, the late Dr. Cleopa Msuya. It always feels great to have you in my life. I dedicate this to you as a symbol of togetherness and infinite love.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

AG_m	Adjacent Geometric Mean
ANN	Artificial Neural Network
AUC	Area Under the Curve
BEST	Basic Education Statistics in Tanzania
DNN	Deep Neural Network
ESDP	Education Sector Development Plan
ETP	Education Training Policy
FGD	Focus Group Discussion
F_m	F-Measure
G_m	Geometric Mean
HMM	Hidden Markov Model
KNN	K-Nearest Neighbor
KPCA	Kernel Principal Component Analysis
LCE	Link-based Cluster Ensembles
LMS	Learning Management System
LR	Logistic Regression
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
MOOCs	Massive Open On-line Courses
PCA	Principal Component Analysis
PGM	Probabilistic Graphical Model
PORALG	President's Office Regional Administration and Local Government
RF	Random Forest
ROS	Random Over Sampling
RTS	Round Table Survey
RUS	Random Under Sampling
SMOTE	Synthetic Minority Over-Sampling Technique
SMOTE ENN	Synthetic Minority Over-Sampling Technique with Edited Nearest Neighbor
SMOTE TOMEK	Synthetic Minority Over-Sampling Technique with Tomek-links
RMSE	Root Mean Square Error

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Student dropout is among the serious challenges that face many schools globally. The student is regarded to have dropped out of school when he or she abandoned the school without finishing studies and be issued with an official certificate of completion (Estêvão & Álvares, 2014; Krstic *et al.*, 2017). A growing body of literature indicates high rates of students' dropout particularly in developing countries as compared to developed countries (Latif *et al.*, 2015; UNESCO, 2017), in girls than boys (Shahidul & Karim, 2015) and in lower secondary as compared to higher level (Hunt *et al.*, 2017). The problem of student dropout is attributed by a myriad of reasons such as economic factors (e.g. parental investment), house-hold level factors (e.g. female involvement in house-hold chores), school-level factors (e.g. school distance) and cultural factors (e.g. early pregnancy) (Shahidul & Karim, 2015). Addressing student dropout problem requires a thorough understanding of the fundamental issues that cause the problem (Ameri *et al.*, 2016).

To this end, several policies, initiatives and strategies have been developed in most parts of the world, particularly in developing countries to address the student dropout problem. Those initiatives include ensuring that all children of school age attain free and quality education, constructing and renovating education facilities (classrooms, latrines, laboratories, playing grounds and dormitories) that are gender and disability friendly, increasing number of qualified teachers, particularly in the field of science and mathematics, reviewing and amends education policies, acts and curriculum that are in line with current pace of science and technology (UNESCO, 2011). Despite these initiatives, the proportion of students dropping out of school is significant and poses a big challenge (Krstic *et al.*, 2017; Otieno, 2016).

Several studies have been conducted on addressing the issue of student dropout by identifying the factors that cause the dropout problem (Lockett & Cornelious, 2015; Murray, 2014; Willging & Johnson, 2009) and examining the techniques and strategies for reducing or eliminating the student dropout problem (Dockery, 2012; Moore, 2017; Rumberger *et al*., 2017; Rutakinikwa, 2016). However, most of these studies focused on the social aspect of student dropout and used methods such as focus group discussions and household surveys in identifying factors perceived to contribute to the risk of student dropout (Baker, 2011; Oruko

et al., 2015), what students perceive to be the consequences of dropout (Amadi *et al.*, 2013; Zheng *et al*., 2015) and the factors that influence and determines whether a student will return to school after dropout (Barrat *et al.*, 2012; Nakpodia, 2010). Although these studies have made an important contribution to the way that student dropout is perceived by the community and students themselves, and have gone further to identify matters for intervention, they have failed to propose workable solutions that will enable the relevant authorities to identify students at risk and intervene prior to the student dropping out of school (Neild *et al.*, 2007).

The success of the above-mentioned initiatives depends on the ability to accurately identify and predict students at risk of dropping out for early intervention. Currently, schools in developing countries are generating tons of data on student attendance, which also includes data on absenteeism, truancy and dropout, however, these data are mostly available in non-digital format and are mainly used during planning and resource allocations (Development Education Research Centre, 2018; Latif *et al.*, 2015). These generated data from schools need to be publicly available in open access platforms for researchers in developing countries to make use of emerging fields such as machine learning in trying to address the student dropout problem in Tanzania (Mduma *et al.*, 2019). Nevertheless, the costs and time for digital data collection and storage has been reported as the biggest challenge in most developing countries (Kim, 2019).

Various studies have been conducted in developed countries in creating data driven predictive approach using machine learning techniques and prove to accurately predict school dropout. However, most of the studies identified factors that are quite different from those identified by the studies in developing countries (Mgala & Mbogho, 2015) and some did not go further in revealing the root causes of the problem (Allensworth & Easton, 2007; Kotsiantis *et al.*, 2003; Neild *et al.*, 2007). The prediction reports from data driven approach systems may alert and help education stakeholders to form an initial hypothesis about the needs of particular students and schools, however, these data driven approaches need to consider and use data from the local context for accurate and better performance (Bowers *et al.*, 2013; Rumberger & Lim, 2008). Thus, there is need to develop a machine learning predictive algorithm using local datasets that can effectively help the relevant authorities such as schools, education officers and the local government on addressing the student dropout problem in secondary schools in Tanzania. By doing so, will accurately identify and predict students at risk of dropping out of

schools for early intervention and policy reviews as an alternative to traditional approaches which in most cases wait until the student have actually dropped out of school.

1.2 Statement of the Problem

Student dropout in secondary schools in many parts of the world is a big challenge to both an individual student and the community in general (Sara *et al.*, 2015). The student dropout rate in Tanzania is 36% in general (UNESCO, 2017). Moreover, 30% of girls in the country are reported to dropout before reaching form 4 (BEST, 2015). In 2015, the government took a crucial step to increase the number of secondary schools and waived all school fees and monetary contributions in public schools as a strategy and an important action to its ambitious education goal and serious commitment of ensuring every child of school age who passed primary school examination gets the right to free education (Human Right Watch, 2017; Ministry of Education, 2015). Furthermore, Education Policies and National Education Act were reviewed and amended to support secondary schools' education, particularly to young girls by imposing heavy punishments to whoever impregnate or married a student thereby causing her to dropout (Human Right Watch, 2017). Despite all these efforts, student dropout is still a problem and the dropout trend in the country have increased from 61 484 students in 2015 to 65 700 students in 2017 (BEST, 2018). This has attributed to the complexity of the student dropout problem which calls for new approaches that will early identify students at risk of dropping out of school for intensive and continuous intervention.

The use of data driven approaches such as machine learning models has gained much attention in other parts of the world, particularly developed countries when addressing society's problems in different sectors such as healthcare, business, industrial and education (Afolabi *et al.*, 2018; Hussain *et al.*, 2019; Panch *et al.*, 2018; Vital Wave Consulting, 2009; Wuest *et al.*, 2016). This is attributed by the fact that, machine learning models when accurately trained, provide convenient and reliable results as compared to the traditional approaches which in most cases have to be implemented at the end of the course i.e. when the student have dropped out of school (Ameri *et al.*, 2016; Lakkaraju *et al.*, 2015; Neild *et al*., 2007).

There are substantial amount of literatures on how different machine learning techniques such as decision tree (Lakkaraju *et al.*, 2015), artificial neural networks, matrix factorization (Elbadrawy *et al.*, 2016; Xu *et al.*, 2017), probabilistic graphical models (Fei & Yeung, 2015) and survival analysis (Ameri *et al.*, 2016) can be applied to develop predictive algorithms for

the students dropout problem using various platforms such as Massive Open Online Course (MOOC) (Chen *et al*., 2012; Liang *et al*., 2016; Fei & Yeung, 2015; Prieto *et al*., 2017) and other Learning Management System (LMS) such as Moodle (Elbadrawy *et al*., 2016; Hung *et al*., 2017; Santana *et al*., 2015). However, most of these works were carried out in developed countries using developed countries datasets and ends up in model development which requires a user to at least have a basic knowledge of machine learning to easily interact with the developed approach (Elbadrawy *et al.*, 2016; Fei & Yeung, 2015; Xu *et al.*, 2017).

1.3 Rationale of the Study

Tanzania and other developing countries need to successfully address the student dropout problem in order to achieve sustainable development. The use of data driven approaches that are trained using local datasets and designed by taking into considerations the end-user interactions are urgently needed for early identification and prediction of students at risk of dropping out, to help education stakeholders and policy makers to make informative decisions and early intervention programs (Márquez-Vera *et al.*, 2016). These approaches will assist in establishing the extent to which different factors contribute to the student dropout problem, allow education stakeholders with no prior knowledge on machine learning easily interacts with, and more importantly, will accurately make use of tons of data on student dropout generated by schools and other institutions on identifying and predicting students at risk of dropping out of school for early intervention.

Therefore, this study attempted to develop data driven approach using datasets from developing countries to help identify and predict students at risk of dropping out of secondary schools in Tanzania in order to facilitate an intervention program to be delivered as an active measure before the student drops out of school.

1.4 Objectives

1.4.1 Main Objective

To develop a data driven approach for identifying and predicting student dropout in secondary schools in Tanzania.

1.4.2 Specific Objectives

(i) To assess machine learning approaches and techniques for student dropout prediction.

(ii) To analyze features that contribute significantly to the student dropout problem.

(iii) To evaluate data balancing techniques for student dropout prediction.

(iv) To develop data driven predictive model for student dropout.

(v) To develop a student dropout predictive prototype and evaluate performance of the developed model.

1.5 Research Questions

(i) What are existing machine learning approaches and techniques for student dropout prediction?

(ii) Which features contribute significantly to the student dropout problem?

(iii) Which is the best data balancing technique for the student dropout prediction?

(iv) How data driven predictive model for student dropout can be developed?

(v) How the student dropout predictive prototype can be developed and what values can it add to the developed model?

1.6 Significance of the Study

The major contribution of this study is the development of the data driven approach to help relevant authorities and other stakeholders in the education sector to make informed decisions and early interventions in addressing the issue of student dropout problem in secondary schools in Tanzania.

Furthermore, this study has identified datasets that will facilitate and promote research activities in the application of machine learning in the education sector. In the like manner, the study also contributed to the understanding on the importance of machine learning models in identifying non-linear factors associated with the problem at hand, in this case students' dropout.

Moreover, the study provides a better theoretical understanding and practical application of machine learning in education sector particularly on addressing student dropout problem.

Also, the developed approach provides information that will help stakeholders such as teachers and parents to gauge the student and school progress by evaluating the schools and individual student dropout rates/status.

Finally, the generated reports from the approach can be used to facilitate planning and budgeting of school requirements based on the number of students by taking into considerations the dropout prediction status.

1.7 Delineation of the Study

This study focused on identifying and predicting students who are at risk of dropping out of secondary schools using datasets from Tanzania. Thus, the study didn't consider ranking students according to their probability of dropping out or forecasting student's dropout trend in the future.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Student Dropout

Student dropout is one of the biggest challenges in most schools and recently has attracted the attention of many researchers (Ashimolowo *et al.*, 2010; Lekwa & Anyaogu, 2016). The problem of student dropout has been reported globally, except for a few countries like Australia, Japan, Norway and Finland, where dropout rates are very low and negligible (OECD, 2015; Rannveig, 2016; Tabuchi *et al.*, 2018; Virtanen & Tuomo, 2016; Vossensteyn *et al.*, 2015). Student dropout definition differs among studies, but in all cases, it is related to student missing or stopping studies before graduation or be issued with an official certificate of completion (UNESCO, 2017). In Tanzania, for example, when a student is absent from school for 90 days continuously, that student will be expelled or considered to have dropped out of school (United Republic of Tanzania., 2015). It is worth highlighting that absenteeism with consent and/or sick leave should not be confused with dropout. There have been many theories and hypotheses that associate various factors to student dropout, however, most reports categorize and focused on cultural and economic reasons (Morara & Chemwei, 2013; Trevor *et al.*, 2018). On many occasions, researchers have reported various factors as straight forward and underlying causes of the student dropout, however, most of them lack strong scientific justification and failed to address the student dropout problem (Branson *et al.*, 2014; Nielsen, 2016).

Low number of qualified teachers have been mentioned as one of the causes of student dropout in most schools, particularly in developing countries (Rumberger & Lim, 2008; Fall & Roberts, 2012). The fewer number of qualified teachers to student ratio has been reported to compromise the school's ability to actively engage students in academic activities and eventually leads to student dropout (Tufi *et al*., 2015; UNICEF, 2017). Furthermore, Kim and Kim (2018) described physical conditions of school facilities such as classrooms and latrines as another reason for student dropout in developing countries. Poor academic performance has also been linked with student dropping out of schools. Krstic *et al*. (2017) in his study to evaluate the student dropout problem in both primary and secondary schools in Serbia, it was found out that, students with higher academic performance are less likely to leave school compared to students with poor academic performance. He further pointed out factors such as insufficient

parents' engagement in a child's schooling and parents or elder siblings' history of dropping out of school can significantly increase the risk of other children becoming early school leavers as well. Moreover, Can *et al*. (2017) and Farah and Upadhyay (2017) identified poverty as among the factors that lead to student dropout, and their studies concluded that students from the poorest and most disadvantaged rural areas tend to have lower educational attainment as compared to the students from higher income families which are most likely to have aspirations that promote persistence. On the other hand, other factors such as positive relationships with peers, absence of violence or bullying, participation in extracurricular activities and different kinds of dialogue in the classroom and school were reported to lower incidence of students dropping out of school (Simic & Krstic, 2017). The student dropout problem requires joint actions among the key players in order to stop. The identification of contributing factors is very important when addressing this problem, thus there is an apparent need for conducting further studies that will focus on causative factors, particularly hidden and non-linear which in one way or another contributes significantly to the student dropout problem.

2.2 Student Dropout Trend in Tanzania

The population structure of Tanzania is predominantly young children (under 15 years) which represents about 44.1% of the entire population (NBS, 2019). This makes education sector a national priority and a key role in the development since a big portion of the country's population constitutes of children of school age (Kassam, 2000). The education system in the country starts with pre-primary education (2 years), followed by primary (7 years), secondary ordinary (4 years), secondary advanced (2 years), and the university level education (3-5 years). (Ministry of Education, 2015). The primary education is mandatory, and the number of children enrolled has been increasing drastically from 67.35% in year 2000 to 94.17% in year 2018 (BEST, 2018). Over the years, the government has worked hard to increase the access and improve the quality of education, capacity building to education stakeholders, and budget increase and direct funding to schools (Ministry of Education, 2015). Despite the above-mentioned efforts, education in most rural parts of Tanzania ended for many children after primary school, and only three out of five children, or 52% of the eligible school population, were enrolled in lower-secondary education and fewer complete secondary education (UNESCO, 2017). To improve this situation, in 2015 the government abolished all school fees and financial contributions to all public secondary schools. The free secondary education helps most children from the lower-income households to continue with secondary school, and this

was reflected by the significant decline of dropout trend (from 94 986 in 2013 to 63 903 in 2016), however the dropout trend was reported to increase again and by 2017 it was 65 700 as shown in Fig. 1 (BEST, 2018). The student dropout is a serious problem which hinders the development of the education sector and government efforts in providing quality education to all children of school age (Mosha, 2014). The benefits of quality education are undoubtedly huge and can lift families and communities out of poverty and increase a country's economic growth. Furthermore, education has been shown to strongly increase individuals' chances of getting employment and eventually have a better life (Human Right Watch, 2017). Finding and implementing solutions to dropout problem will have implications well beyond the benefits to individual students. Moreover, enabling students to complete their education means investing in the future progress and better standards of life. To make efforts that will improve the situation, a clear understanding of the extent, causes, consequences, and policy responses to student dropout is required.

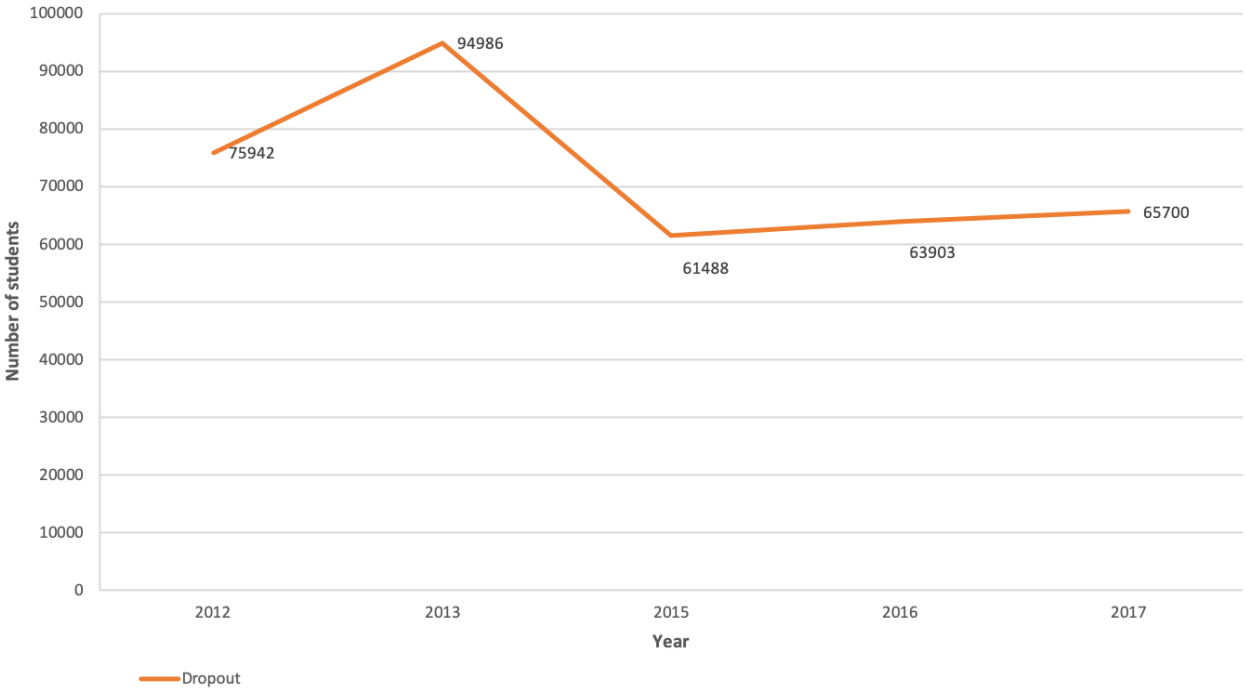


Figure 1: Students' dropout trend in Tanzania (2012-2017)

2.3 Machine Learning in Education

Over the past two decades, there has been a significant advancement in machine learning in healthcare, engineering, agriculture, finance and most recently education sector. This field

emerged as the method of choice for developing practical software, by train a model using datasets for various applications (Jordan & Mitchell, 2015). There are several areas where education sector can take advantage of this emerging field of machine learning to develop workable solutions that will positively impact the communities. The study conducted by Center for Digital Technology and Management (2015) reported on increased use of machine learning in education, due to the rise in the amount of education data available. This advancement made it possible for several studies to apply machine learning techniques in improving educational quality through assessing the learning quality (Ciolacu *et al.*, 2017), knowledge areas related to learning and content analytics (Waters *et al.*, 2014), knowledge tracing (Yudelson *et al.*, 2013), learning material enhancement (Rakesh *et al.*, 2014) and student dropout prediction (Beck & Davidson, 2001). The use of machine learning techniques for educational purpose show promising results, hence schools and other institutions can use the developed models to discover meaningful patterns for solving everyday challenges.

Furthermore, Kučak *et al.* (2018) conducted a survey of research trends on the application of machine learning in education and identified several studies which show how reliable machine learning models can be applied to address issues in improving student retention, testing students, grading students and predicting student performance. The study conducted by Morris *et al*. (2005) in predicting student retention in online courses using predictive discriminant analysis achieved an accuracy ratio of 74.5%. Moreover, Herzog (2006) estimated student retention at the higher learning institutions using neural networks, rule induction and multinomial logistic regression and all three methods achieved similar predictive accuracy of 75% and 84% in the mid and end of the year respectively. Additionally, the study conducted by Chung and Lee (2019) in predicting students' dropout in high schools in Korea using random forest gave excellent predictive accuracy of 95%. Despite the good predictive accuracy of various machine learning models in the education sector, fewer studies have been conducted on how predictive models can be used in identifying the reasons behind the student dropout problem and predicting students at risk of dropping out of schools in developing countries.

2.4 Identification of Factors for Student Dropout

Student dropout is linked to various factors. It is necessary to thoroughly understand each factor and its contributions when analyzing the root causes of dropout (Stempel *et al.*, 2017). The key characteristic is the existence of a relationship between the factor and student dropout. Identifying the most probable factors is close to eliminating the dropout problem. The most

common method of identifying factors contributing to student dropout is prior knowledge based on the perceptions, experience or beliefs (Habibipour *et al.*, 2018). Several studies have reported the application of this traditional approach in different parts of the world prior to the use of the emerging field of machine learning in the identification of factors for student dropout in education (Hailikari *et al.*, 2007; Rittle-Johnson *et al.*, 2009; Archambault *et al.*, 2009; Chen, 2012). The prior knowledge is very useful and time effective when working with an existing limited set of factors and there is no room for any maneuvers (Hailikari *et al.*, 2007). However, the use of this approach alone can potentially lead to biasness and restrict the identification of potential non-linear contributors of the problem.

Despite the popularity and wide use of traditional approaches in identifying factors that lead to student dropout particularly in the developing countries, fewer studies have been reported to apply the emerging field of Machine Learning (ML) in student dropout. Machine learning techniques such as variable ranking, a permutation of the feature engineering, search and embedded method has been applied in identifying features that contribute significantly to the student dropout problem (Bouaguel *et al*., 2015; Mduma *et al*., 2019; Ma *et al*., 2017). The application of these techniques has been preferred due to their simplicity, scalability, and good empirical success (Sun *et al*., 2017). However, the successful application of aforementioned approaches requires a knowledge in ML, which poses a challenge in the developing world, particularly in the education sector where most of stakeholders lack basic knowledge in computing (Mgala, 2016; Lakkaraju *et al.*, 2015), thus the use of multi approaches during identification of factors and development of simple and user friendly models with considerations of end users is highly recommended.

2.5 Data for Model Development

Machine learning requires huge amount of data to train the model (Mullainathan & Spiess, 2017). In order for a machine learning model to perform accurately, data balances is essential (Krawczyk, 2016; López *et al.*, 2013). In the real world, many datasets are not well balanced whereby one class is under-represented (minority) than the other (majority) (Abdi & Hashemi, 2014; Borowska & Topczewska, 2016; Galar *et al.*, 2016; Krawczyk, 2016; Lin & Chen, 2012; Mazumder *et al.*, 2015). In the education context, classification of imbalanced problem is being recognized in the student dropout field, because the number of registered students is larger than the number of dropout students (Thammasiri *et al.*, 2014). According to Gao (2015), the

imbalanced ratio is about at least 1:10 and in most cases the minority class usually represents the target group (López *et al.*, 2013).

On handling imbalance datasets, several approaches such as algorithmic modification and cost-sensitive learning have been applied (Elhassan *et al.*, 2016; Hoens & Chawla, 2013). Algorithmic modification focuses on changing the learning algorithm to adapt the imbalance data setting and works with the algorithmic level (Elhassan *et al.*, 2016), while cost-sensitive learning approach takes costs into consideration with the aim of minimizing the costs associated with the learning process (Shilbayeh, 2015). Several studies have reported the limitations of the two approaches in terms of cost and time when handling data sets with highly skewed data as in the case of student dropout, and proposed the use of other approaches or reduce the data sets for the effective learning process (Márquez-Vera *et al.*, 2016; Neill *et al.*, 2020; Weiss *et al.*, 2007). However, reducing data sets might exclude important samples which may affect the learning process, hence the use of other approaches for handling highly skewed data sets is recommended.

2.6 Machine Learning Approaches and Student Dropout

Machine learning approaches are efficient and capable of finding solutions to several linear and non-linear problems such as plant control, forecasting, prediction, robotics and so many others (Sathya & Abraham, 2013). Regression and classification are among machine learning categorized under supervised learning due to their concept of utilizing known datasets to make predictions. The difference between these approaches is based on the output variable which are numerical for regression and categorical for classification. In addressing the problem of student dropout, several studies have applied different ML approaches (Table 1) such Artificial Neural Network (ANN), Decision Tree, Naive Bayesian Algorithm, Association Rules Mining, Based Algorithm, Cox Regression, Logistic Regression, Random Forest and Simple Logistic among others (Kumar *et al.*, 2017b).

The study conducted by Aguiar *et al.* (2015) highlighted students at risk of not graduating on time in the United States of America (USA) using ML approaches such as Cox Regression, Logistic Regression and Random Forest. The model was trained using dataset collected from the school district, and the performance of the model was evaluated using Accuracy, Mean Absolute Error (MAE), Information Gain (IG), Gini Impurity (GI), Stepwise Regression (SR), Single Feature Performance (SFP1) and Random Forest (RF). The results showed that RF

achieved better performance compared to other ML approaches. On the other hand, Halland *et al.* (2015) conducted a study to address similar problem of student dropout but in secondary school using datasets collected from Danish high schools. The study used Support Vector Machines (SVMs) with Gaussian Kernels, Random Forests CART and Naive Bayes to build the model. The model was evaluated using Accuracy and Area Under the ROC Curve (AUC) and again the results showed that, Random Forest model achieved better performance compared to others. However, the issue of data imbalance was ignored in both studies, and needs to be addressed for better performance and predictive accuracy of the model.

Furthermore, Prieto *et al.* (2017) conducted a study for student dropout prediction in Massive Open On-line Course (MOOC). Logistic Regression and Feed-forward Neural Networks were used to build the model, while the performance of the model was evaluated using Area Under the ROC Curve (AUC). The results revealed that, Feed-forward Neural Network outperformed other approaches. Ameri *et al.* (2016) on the other hand, developed a survival analysis framework with the aim of identifying at-risk students using Cox proportional hazards model (Cox) and applied time dependent Cox (TD-Cox). The study used a dataset of students enrolled at Wayne State University (WSU). Accuracy, F-measure (F_m), Area Under the Curve (AUC) and Mean Absolute Error (MAE) were used to evaluate the performance of the model and Cox-based framework achieved the best results compared to other methods. However, the studies didn't address the issue of results interpretation, therefore interaction with the developed models required a prior knowledge on ML.

Despite the presence of various ML approaches for model development, Artificial Neural Network and Decision Tree have been reported by many researchers as the most commonly used approaches in the field of education particularly in student dropout predictions (Joseph, 2014; Shahiri *et al.*, 2015). Additionally, the Neural Network has been reported to offer more benefits over other approaches due to its ability to detect all possible interactions between features (Gray *et al.*, 2014) and perform complex nonlinear relationship between dependent and independent variables (Arsad *et al.*, 2013). The Decision Tree on the other hand, has been reported to provide less information on the relationship between the predictors and the response, hence making it a less preferred approach (Natek & Zwilling, 2014).

There are several machine learning approaches for developing models, however the choice of the approach hugely depends on the nature of the data (labeled or unlabeled) and performance of the model as reported by other researchers. Knowledge on the data plays a key role in choosing the right algorithm for the problem at hand due to the fact that some algorithms can work with smaller sample while others require tons of data.

2.7 Evaluation Measures for Student Dropout

In assessing the performance of machine learning models, one of the key factors is the evaluation criteria (Kumar *et al*., 2017). The selection of appropriate measure is highly determined by the nature of the problem (classification or regression) and/or the nature of the dataset (balance or imbalance). Several studies have applied different evaluation metrics on addressing the problem of student dropout (Table 2), however, the most commonly used metrics are Accuracy, Area Under the Curve (AUC), Mean Squared Error (MSE) among others (Santana *et al*., 2015; Xu *et al*., 2017; Johnson *et al*., 2015). Accuracy as a statistical measure for quantifying the degree of correctness has the ability to give the precise predictive results when the number of samples belonging to each class are equally distributed (Ameri *et al.*, 2016; Lakkaraju *et al.*, 2015; Rovira *et al*., 2017). Area Under the Curve (AUC) on the other hand, is used in binary classification problem to evaluate the probability that the model will rank a randomly chosen positive sample higher than a randomly chosen negative sample (Fei & Yeung, 2015; Halland *et al*., 2015; Liang *et al*., 2016; Prieto *et al.,* 2017). On the contrary, the Mean Squared Error (MSE) is mostly used in the regression problem and can easily compute the gradient by taking the average of the square of the difference between the original values and the predicted values (Elbadrawy *et al*., 2016; Iam-On & Boongoen, 2017). Despite the ability of these metrics for evaluating performance of the ML models, other studies have reported their limitations in terms of misinterpretations of the results and effects on the minority classes (Liang *et al*., 2016; Lin & Chen, 2012; López *et al*., 2013; Longadge *et al*., 2013) hence, the application of several metrics is highly recommended when evaluating the performance of ML models.

Table 1: Summary of the various ML approaches used in model development

Source	Technique	Performance	Limitation
Aguiar *et al.* (2015)	Cox Regression, Logistic Regression Model and Random Forest	Random Forest	Data imbalance
Halland *et al.* (2015)	Support Vector Machines with Gaussian Kernels, Random Forest CART and Naive Bayes Classifier	Random Forest	
Rovira *et al.* (2017)	Logistic Regression, Gaussian Naive Bayes, Support Vector Machines, Random Forest and Adaptive Boosting	Random Forest and Adaptive Boost	
Mgala and Mbogho (2015)	Logistic Regression, Multilayer Perceptron, Sequential Minimal Optimization Algorithm Bayesian Network Classifiers, Naïve Bayes Classifier, Lazy Learners, Random Forest.	Logistic Regression	
Elbadrawy *et al*. (2016)	Personalized Linear Multi Regression, Matrix Factorization, Random Forest, Mean of Means and Uniform Random Guessing	Personalized Linear Multi Regression and Matrix Factorization	
Prieto *et al.* (2017)	Logistic Regression, Feed forward Neural Networks	Feed-forward Neural Network	Result interpretations
Ameri *et al*. (2016)	Cox Proportional Hazards and Time-dependent Cox	Cox-based Framework	
Iam-On and Boongoen (2017)	Data Transformation Model and Cluster ensembles	Data Transformation Model	

Table 2: Summary of the various metrics used in model evaluation

Source	Problem	Metrics
M´arquez-Vera *et al*. (2016)	Mining best rule to predict student dropout	Geometric Mean
Liang *et al*. (2016)	Developing dropout predictive model	Area Under the Curve (AUC)
Poh and Smythe (2015)	Predicting student performance	The error residuals
Fei and Yeung (2015)	Predicting dropout in Massive Open Online Courses (MOOCs)	Area Under the Curve (AUC)
Hung *et al*. (2017)	Identifying at-risk students in online program	Accuracy and misclassification rates
Johnson *et al*. (2015)	Identifying students at risk of not graduating on time	Precision
Xu *et al*. (2017)	Tracking and predicting student performance	Mean square errors
Santana *et al*. (2015)	Identifying students with dropout profiles	Accuracy and a false positive rate

CHAPTER THREE

MATERIALS AND METHODS

3.1 Materials

3.1.1 Description of the Study Area

The data on factors contributing to student dropout were collected from five districts namely Mbeya, Nzega, Rufiji, Kisarawe and Arusha in Tanzania (Fig. 2). The selection of mentioned districts was done out of consideration for dropout prevalence (districts with high, medium and low) and geographical representation (coast, southern, north and west). Furthermore, data from Tanzania (government and non-governmental organization) and India, which are publicly available were used in model development.

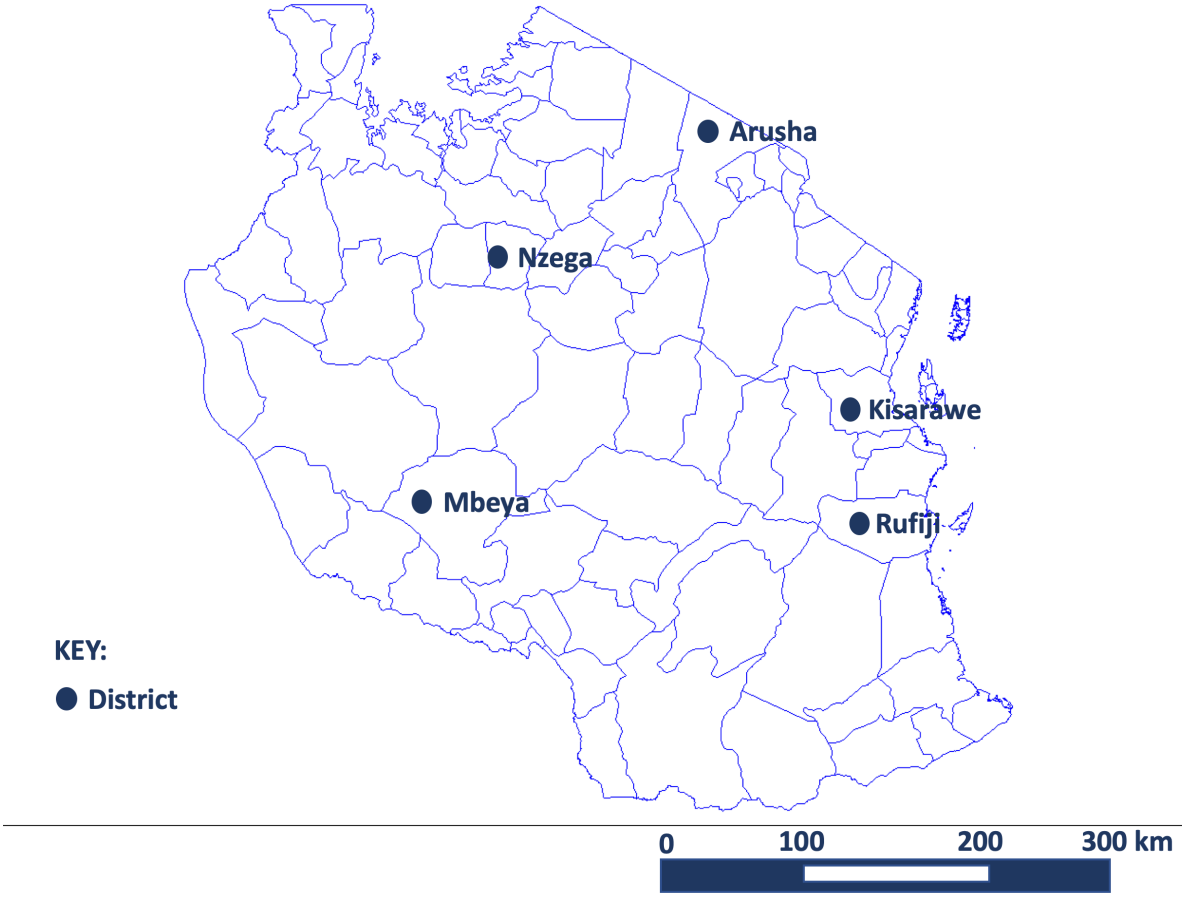


Figure 2: Map of Tanzania showing the study area (Arusha, Kisarawe, Mbeya, Nzega and Rufiji)

3.1.2 Datasets used in Model Development

The study used three different datasets. The first dataset was Uwezo data on learning at the country level in Tanzania which was collected by Twaweza (non-governmental organization in 2015 with the aim of assessing children's learning levels across hundreds of thousands of households. The dataset is publicly available (https://www.twaweza.org/go/uwezo-datasets) and contains approximately 110356 samples and 20 features on student dropout which were our target variable (Fig. 3). The second dataset was from India, which was collected by different governmental and non-governmental organization in 2016 to assess the student dropout in the country. This dataset is also publicly available (https://www.kaggle.com/imrandude/studentdropindia2016) and contains 19101 samples with 15 features (Fig. 4). The third dataset was school level dataset collected by the President's Office Regional Administration and Local Government (PORALG). This dataset was integrated with the publicly available data from the government open data portal (http://opendata.go.tz/dataset) and contains 145 samples with 21 features (Fig. 5). This dataset was used to support the visualization of the school dropout trend. Selection of these three datasets took into consideration of data from developing countries (Tanzania and India) and to ensure reliability the data sets were collected from the government, reputable non-governmental organization, and the world's largest data science community.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		id_districtn	id_regionna	id_village	villmtaanan	hh_size	mealsperda	Student_age	Dropout	SchoolhasG	PTR	PCR	GPLR	BPLR	PTMR	Student_rea	Parentteach	ParentChecl	EA_area	Studentsex	Income__source	
2	0	Kondoa	Dodoma	1	Bumbuta	5	3	8	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	2	2	
3	1	Kondoa	Dodoma	1	Bumbuta	8	3	10	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	1	2	
4	2	Kondoa	Dodoma	1	Bumbuta	8	2	10	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	2	2	
5	3	Kondoa	Dodoma	1	Bumbuta	8	3	7	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	1	2	
6	4	Kondoa	Dodoma	1	Bumbuta	8	2	13	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	1	2	
7	5	Kondoa	Dodoma	1	Bumbuta	6	3	10	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	1	1	2	2	2	
8	6	Kondoa	Dodoma	1	Bumbuta	6	3	14	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	1	2	
9	7	Kondoa	Dodoma	1	Bumbuta	6	3	12	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	2	2	2	2	
10	8	Kondoa	Dodoma	1	Bumbuta	6	3	8	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	1	2	
11	9	Kondoa	Dodoma	1	Bumbuta	6	3	15	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	1	1	2	2	2	
12	10	Kondoa	Dodoma	1	Bumbuta	5	3	11	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	2	2	
13	11	Kondoa	Dodoma	1	Bumbuta	5	3	7	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	1	2	1	2	
14	12	Kondoa	Dodoma	1	Bumbuta	6	3	7	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	2	2	1	2	
15	13	Kondoa	Dodoma	1	Bumbuta	4	3	10	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	2	2	
16	14	Kondoa	Dodoma	1	Bumbuta	4	3	12	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	2	2	
17	15	Kondoa	Dodoma	1	Bumbuta	3	2	16	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	1	1	2	1	2	
18	16	Kondoa	Dodoma	1	Bumbuta	3	2	11	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	1	1	2	1	2	
19	17	Kondoa	Dodoma	1	Bumbuta	5	3	9	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	2	2	
20	18	Kondoa	Dodoma	1	Bumbuta	6	3	11	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	1	1	2	1	2	
21	19	Kondoa	Dodoma	1	Bumbuta	6	3	13	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	1	1	2	1	2	
22	20	Kondoa	Dodoma	1	Bumbuta	6	3	7	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	2	2	2	2	
23	21	Kondoa	Dodoma	1	Bumbuta	6	3	14	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	2	2	1	2	
24	22	Kondoa	Dodoma	1	Bumbuta	4	3	7	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	1	2	
25	23	Kondoa	Dodoma	1	Bumbuta	6	3	11	1	2	47.875	42.55556	66.33334	61.33333	29.46154	2	2	2	2	2	2	
26	24	Kondoa	Dodoma	1	Bumbuta	5	3	12	1	2	47.875	42.55556	66.33334	61.33333	29.46154	1	2	1	2	1	2	
27	25	Kondoa	Dodoma	2	Pahi	7	2	9	1	2	26.14286	40.66667	35	38.2	183	1	1	2	2	1	2	
28	26	Kondoa	Dodoma	2	Pahi	4	3	11	1	2	26.14286	40.66667	35	38.2	183	1	1	1	2	2	2	
29	27	Kondoa	Dodoma	2	Pahi	12	2	15	1	2	26.14286	40.66667	35	38.2	183	1	2	1	2	2	2	
30	28	Kondoa	Dodoma	2	Pahi	4	2	10	1	2	26.14286	40.66667	35	38.2	183	1	2	1	2	2	2	
31	29	Kondoa	Dodoma	2	Pahi	5	3	6	1	2	26.14286	40.66667	35	38.2	183	1	2	1	2	2	2	
32	30	Kondoa	Dodoma	2	Pahi	4	3	10	1	2	26.14286	40.66667	35	38.2	183	2	2	2	2	1	2	
33	31	Kondoa	Dodoma	2	Pahi	5	3	12	1	2	26.14286	40.66667	35	38.2	183	1	2	1	2	2	2	
34	32	Kondoa	Dodoma	2	Pahi	4	3	7	1	2	26.14286	40.66667	35	38.2	183	2	2	2	2	1	2	
35	33	Kondoa	Dodoma	2	Pahi	4	2	10	1	2	26.14286	40.66667	35	38.2	183	1	2	1	2	2	2	
36	34	Kondoa	Dodoma	2	Pahi	3	3	14	1	2	26.14286	40.66667	35	38.2	183	1	1	1	2	1	2	

Figure 3: The original dataset before preprocessing - Uwezo

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	continue_dr	student_id	gender	caste	mathematic	english_mar	science_mar	science_tea	languages_t	guardian	internet	school_id	total_studen	total_toilets	establishment_year	
2	continue	s00001	F	SC	0.409	0.514	0.409	6	0	mother	TRUE	310	262	28	1985	
3	continue	s00002	F	BC	0.29	0.512	0.29	4	7	mother	TRUE	328	356	14	1943	
4	continue	s00003	F	OC	0.602	0.666	0.602	4	2	mother	FALSE	322	179	8	1955	
5	continue	s00004	F	BC	0.378	0.526	0.378	8	7	mother	TRUE	305	354	86	1986	
6	continue	s00005	F	OC	0.536	0.614	0.536	9	4	other	TRUE	360	273	2	1995	
7	continue	s00006	F	BC	0.594	0.519	0.594	4	8	mother	TRUE	333	335	43	1916	
8	continue	s00007	F	OC	0.177	0.525	0.177	6	7	father	TRUE	351	251	100	1924	
9	continue	s00008	F	OC	0.48	0.457	0.48	2	9	mother	TRUE	372	99	14	1956	
10	continue	s00009	F	OC	0.821	0.728	0.821	2	2	father	TRUE	389	299	1	1921	
11	continue	s00010	F	BC	0.418	0.322	0.418	8	10	father	FALSE	342	295	18	1907	
12	continue	s00011	F	OC	0.757	0.664	0.757	8	6	mother	TRUE	324	170	31	1949	
13	continue	s00012	F	OC	0.133	0.566	0.133	2	2	mother	TRUE	338	145	2	1929	
14	continue	s00013	F	SC	0.566	0.495	0.566	0	4	father	TRUE	366	62	28	1897	
15	continue	s00014	F	BC	0.413	0.598	0.413	1	6	father	TRUE	301	49	14	1844	
16	continue	s00015	F	BC	0.461	0.524	0.461	0	3	mother	TRUE	392	469	14	1905	
17	continue	s00016	F	OC	0.742	0.672	0.742	3	12	mother	TRUE	383	132	14	1996	
18	drop	s00017	F	BC	0.503	0.523	0.503	9	0	father	TRUE	362	397	5	1950	
19	continue	s00018	F	OC	0.32	0.408	0.32	9	6	mother	TRUE	300	198	28	1913	
20	continue	s00019	F	SC	0.306	0.485	0.306	5	6	father	FALSE	393	111	10	2006	
21	continue	s00020	F	BC	0.107	0.507	0.107	7	5	father	TRUE	392	469	14	1905	
22	continue	s00021	F	ST	0.38	0.426	0.38	9	4	mother	TRUE	353	359	8	1986	
23	continue	s00022	F	BC	0.746	0.669	0.746	3	6	mother	TRUE	341	430	44	1959	
24	drop	s00023	F	ST	0.681	0.758	0.681	5	3	mother	TRUE	386	220	30	1847	
25	continue	s00024	F	SC	0.895	0.615	0.895	4	0	mother	TRUE	365	444	14	1970	
26	continue	s00025	F	BC	0.132	0.759	0.132	4	0	mother	TRUE	318	54	17	1867	
27	continue	s00026	F	BC	0.456	0.46	0.456	2	12	mother	TRUE	353	359	8	1986	
28	continue	s00027	F	BC	0.444	0.574	0.444	4	0	father	TRUE	382	470	71	1896	
29	continue	s00028	F	SC	0.304	0.556	0.304	9	10	father	TRUE	313	399	14	1916	
30	continue	s00029	F	OC	0.362	0.703	0.362	0	5	mother	TRUE	310	262	28	1985	

Figure 4: The original dataset before preprocessing - India

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	REGION	DISTRICT	WARD	SCHOOL N	NAME TYPE OWNERS	Boys Latrines	Girls Latrines	Classrooms	Dropout Male	Dropout Female	Dropout Total	Total Enroll	Total Teacher	PTR	Qualified Teacher	PQTR	PCR	BPLR	GPLR	Male Enrollment	Female Enrollment
2	ARUSHA	Arusha CC	Baraa	SORENYI	Government	9		16				680	29	23	29	23	42.5	36		324	356
3	ARUSHA	Arusha CC	Baraa	BARAA	Government	8	26	19				838	41	20	40	21	44.10526	53.125	15.8846154	425	413
4	ARUSHA	Arusha CC	Daraja Ii	FELIX MRE	Government	12		17	8	7	15	1363	48	28	47	29	80.17647	54.66667		656	707
5	ARUSHA	Arusha CC	Daraja Ii	ARUSHA GI	Non-Government		10	4				151	18	8	15	10	37.75		15.1	0	151
6	ARUSHA	Arusha CC	Elerai	ELERAI	Government	12	4	20	10	5	15	1421	58	25	58	25	71.05	57.91667	181.5	695	726
7	ARUSHA	Arusha CC	Engutoto	KORONA	Government	4		10	0	3	3	93	16	6	15	6	9.3	10.75		43	50
8	ARUSHA	Arusha CC	Engutoto	NJIRO	Government	6	5	10	3	4	7	426	31	14	31	14	42.6	35.66667	42.4	214	212
9	ARUSHA	Arusha CC	Engutoto	EL-SHAMM	Non-Government	0		4				89	5	18	3	30	22.25				
10	ARUSHA	Arusha CC	Kaloleni	KALOLENI	Government	6	4	16				169	11	15	5	34	10.5625	95.66667	130.5	574	522
11	ARUSHA	Arusha CC	Kati	ARUSHA	Government	10		24	7	15	22	1635	87	19	83	20	68.125	82.9		829	806
12	ARUSHA	Arusha CC	Kati	BONDENI	Non-Government	6		14	10	9	19	152	16	10	14	11	10.85714	16.16667		97	55
13	ARUSHA	Arusha CC	Kimandolu	BRAINY HE	Non-Government	0		8				17	5	3	3	6	2.125			17	0
14	ARUSHA	Arusha CC	Kimandolu	SUYE	Government	3	10	18	3	3	6	918	39	24	39	24	51	155.3333	45.2	466	452
15	ARUSHA	Arusha CC	Kimandolu	KIMASEKI	Government	13		27				1205	53	23	52	23	44.62963	48.92308		636	569
16	ARUSHA	Arusha CC	Kimandolu	KIMANDOI	Non-Government	10	15	8	2	2	4	295	14	21	11	27	36.875	13.4	10.7333333	134	161
17	ARUSHA	Arusha CC	Lemara	RENEA	Non-Government	0		4				121	13	9	8	15	30.25			121	0
18	ARUSHA	Arusha CC	Lemara	NAURA	Government	5	12	13	3	2	5	453	36	13	36	13	34.84615	44.4	19.25	222	231
19	ARUSHA	Arusha CC	Lemara	LEMARA	Government	7	9	28	2	1	3	1158	45	26	43	27	41.35714	78.57143	67.5555556	550	608
20	ARUSHA	Arusha CC	Lemara	NOTREDAN	Non-Government	20	5	5				212	12	18	11	19	42.4	0.3	41.2	6	206
21	ARUSHA	Arusha CC	Levolosi	MARANAT	Non-Government	3	5	4				34	7	5	5	7	8.5	4.333333	4.2	13	21
22	ARUSHA	Arusha CC	Moshono	CHARLES D	Non-Government																
23	ARUSHA	Arusha CC	Moshono	LOSIRWAY	Government	6		8	6	7	13	427	21	20	21	20	53.375	33.33333		200	227
24	ARUSHA	Arusha CC	Moshono	MOSHONC	Government	6	5	13	5	3	8	622	43	14	42	15	47.84615	48.66667	66	292	330
25	ARUSHA	Arusha CC	Moshono	ADILI	Non-Government	0		16				185	11	17	7	26	11.5625			185	0
26	ARUSHA	Arusha CC	Moshono	ARUSHA M	Non-Government	12	5	15				217	21	10	9	24	14.46667	11.16667	16.6	134	83
27	ARUSHA	Arusha CC	Moshono	ST. MONIC	Non-Government		16	8				76	8	10	8	10	9.5		4.75	0	76
28	ARUSHA	Arusha CC	Muriet	MURIET SP	Non-Government	3	6	6				45	7	6	7	6	7.5	9.666667	2.66666667	29	16
29	ARUSHA	Arusha CC	Muriet	HADY	Non-Government	8	16	8				49	7	7	7	7	6.125	2.875	1.625	23	26
30	ARUSHA	Arusha CC	Muriet	KINANA	Government	12		11	0	1	1	629	24	26	24	26	57.18182	20.5		246	383
31	ARUSHA	Arusha CC	Ngarenaro	NGARENAR	Government	8		24	2	6	8	1509	42	36	42	36	62.875	89.125		713	796
32	ARUSHA	Arusha CC	Ngarenaro	ENGARENA	Non-Government	6	11	8				177	10	18	10	18	22.125	10.5	10.3636364	63	114
33	ARUSHA	Arusha CC	Olasiti	ARUSHA CI	Non-Government	0		7				175	17	10	17	10	25			175	0
34	ARUSHA	Arusha CC	Olasiti	ARUSHA GI	Government		10	4				150	10	15	10	15	37.5		15	0	150
35	ARUSHA	Arusha CC	Olasiti	OLASITI	Government	5	11	18	1	0	1	1071	35	31	35	31	59.5	98.2	52.7272727	491	580
36	ARUSHA	Arusha CC	Olasiti	PEACE HOL	Non-Government	22		12	1	0	1	516	22	23	19	27	43	12.54545		276	240
37	ARUSHA	Arusha CC	Olmoti	OLMOTI	Government	4		4	1	0	1	80	8	10	8	10	20	8.25		33	47
38	ARUSHA	Arusha CC	Oloirien	OLORIENI	Government	20		17	5	2	7						0	24.85		497	550
39	ARUSHA	Arusha CC	Osunyai Jr	SOMBETIN	Government	14	16	19	16	12	28	1257	45	28	45	28	66.15789	43	40.9375	602	655
40	ARUSHA	Arusha CC	Sakina	SAKINA	Non-Government	16		20	0	1	1	28	8	4	8	4	1.4	0.25		4	24
41	ARUSHA	Arusha CC	Sekei	PRIME	Non-Government	7	8	4				48	10	5	8	6	12	3.714286	2.75	26	22
42	ARUSHA	Arusha CC	Sinoni	SINONI	Government	12	15	19	42	37	79	1704	50	34	47	36	89.68421	68.66667	58.6666667	824	880

Figure 5: The original dataset before preprocessing - PORALG

3.1.3 Programming Language

The programming language used in this study was Python. The selection of this programming language took into considerations its ability to offer a vast set of open-source libraries to support machine learning.

3.2 Methods

3.2.1 Design Science Research Approach

In carrying out this study, a design science research approach was used. This approach was selected due to its ability to solve a problem with the focus on the creation and investigation of technological artifacts. Since the main objective of this study was to develop data driven approach for predicting student dropout in secondary schools, design science research gives the necessary framework for implementation of the developed study. The study followed an iterative design cycle adopted from Vaishnavi and Kuechler (2015) (Fig. 6).

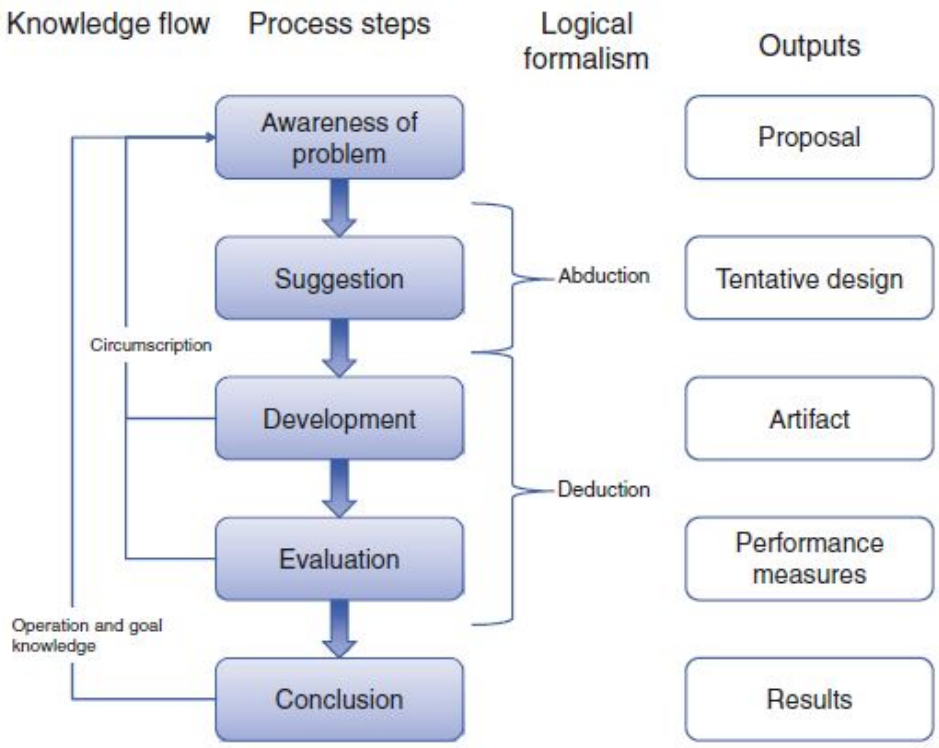


Figure 6: Design Science Research (DSR) model

The DSR model adopted in this study involve five process steps (Fig. 7) which were aligned with the specific research objectives.

(i) Awareness of Problem

The first process step focused on reviewing literatures and conducting focus group discussion (FGD) and round table survey (RTS) with the identified key stakeholders. The purpose of the literature review was to summarize and identifying gaps from the existing research so as to provide better understanding of the research problem, build a theoretical and practical algorithm related to a specific research question and demarcate the scope of the study. The FGD and RTS were conducted with the aim of extracting knowledge and perception regarding student dropout from the key stakeholders.

(ii) Suggestion

In the second process step, literature review was mainly used to facilitates knowledge acquisition. Detailed survey on existing knowledge related to machine learning approaches on student dropout prediction was conducted in academic journals, books, and case studies with the focus on student dropout prediction, student academic performance prediction and student final result prediction. Several databases such as ResearchGate, Elsevier, Association for Computing Machinery (ACM), Science Direct, Springer Link, IEEE Xplore, and other computer science journals were used to search for research articles. Keywords used to search for articles include prediction, machine learning, dropout, technique and approach. The materials taken into consideration were from the year 2013 to 2020 which includes journal articles, conference paper, workshop papers, topics related blogs, expert lectures or talks and reports from research and other organizations.

(iii) Development

The third step focused on the development of the student dropout predictive model. Since the process involves machine learning as the subset of data driven approach, empirical work was conducted to facilitate model development. Empirical work comprised all kinds of data preprocessing. Students related data from different sources were analyzed and integrated as a prior requirement for predictive model development to integrate different attributes from different data sources. All datasets were included in the training and validation of the predictive model.

(iv) Evaluation

In the fourth process step, the application of the developed model for predicting student dropout was considered. The focus was to assess the model's performance on addressing the student dropout problem. Data driven predictive model which was developed and validated was then evaluated using unseen test set in order to observe how the model will behave in a real environment. The developed model was deployed in a prototype to facilitate interpretation of machine learning results. Furthermore, the school-level dataset was used to support visualization of the developed prototype.

(v) Conclusion

Lastly, the outcomes and findings of the entire research were effectively communicated in this process step. The results and findings from this research were communicated to both technical and managerial audiences through journal publications, conferences, workshops, seminars and poster presentations.

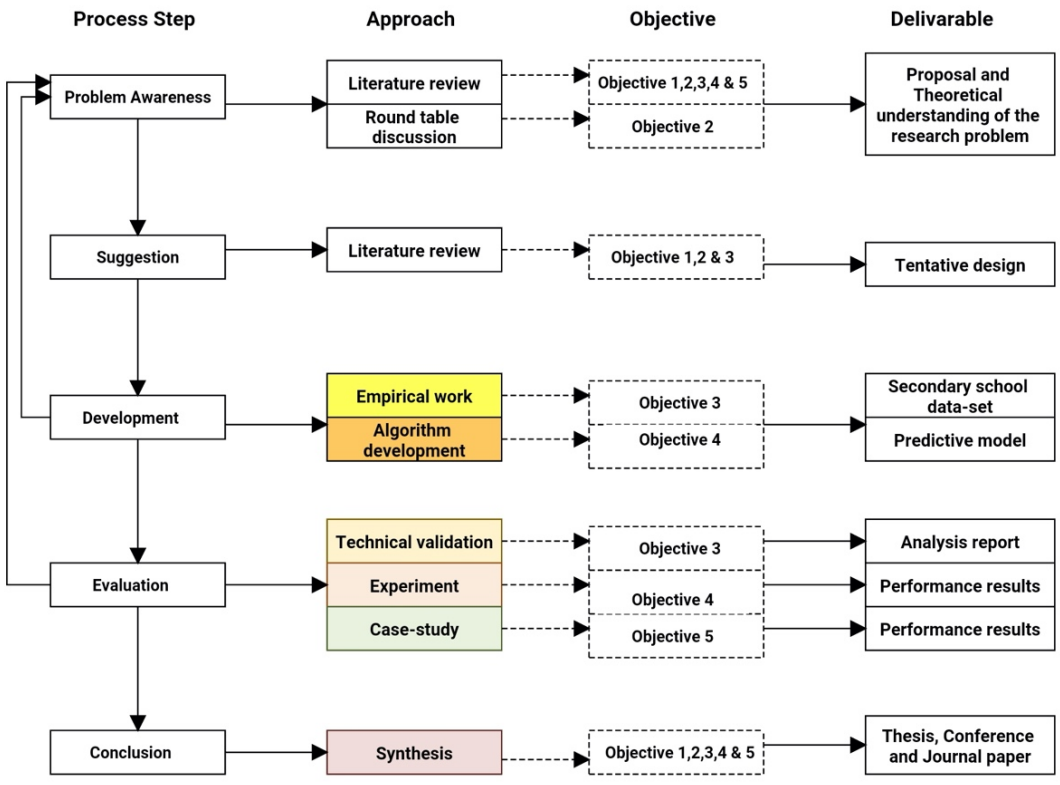


Figure 7: Design science research process

3.2.2 Experimental Procedures for Model Development

The dataset was divided into train (60%), validation (20%) and test (20%) as described by Wu *et al*. (2013). In this experiment, k=5 fold out-of-bag overall cross validation was used and the entire process involves executing all selected classification algorithms in which all executions were repeated 5 times using different train/validation/test partitions of the data set. This cross-validation procedure divided the data set into 5 roughly equal parts. For each part, it trains the model using the four remaining parts and computes the test error by classifying the given part, and the results for the five test partitions were averaged. The overall experimental procedures is summarized in Fig. 8

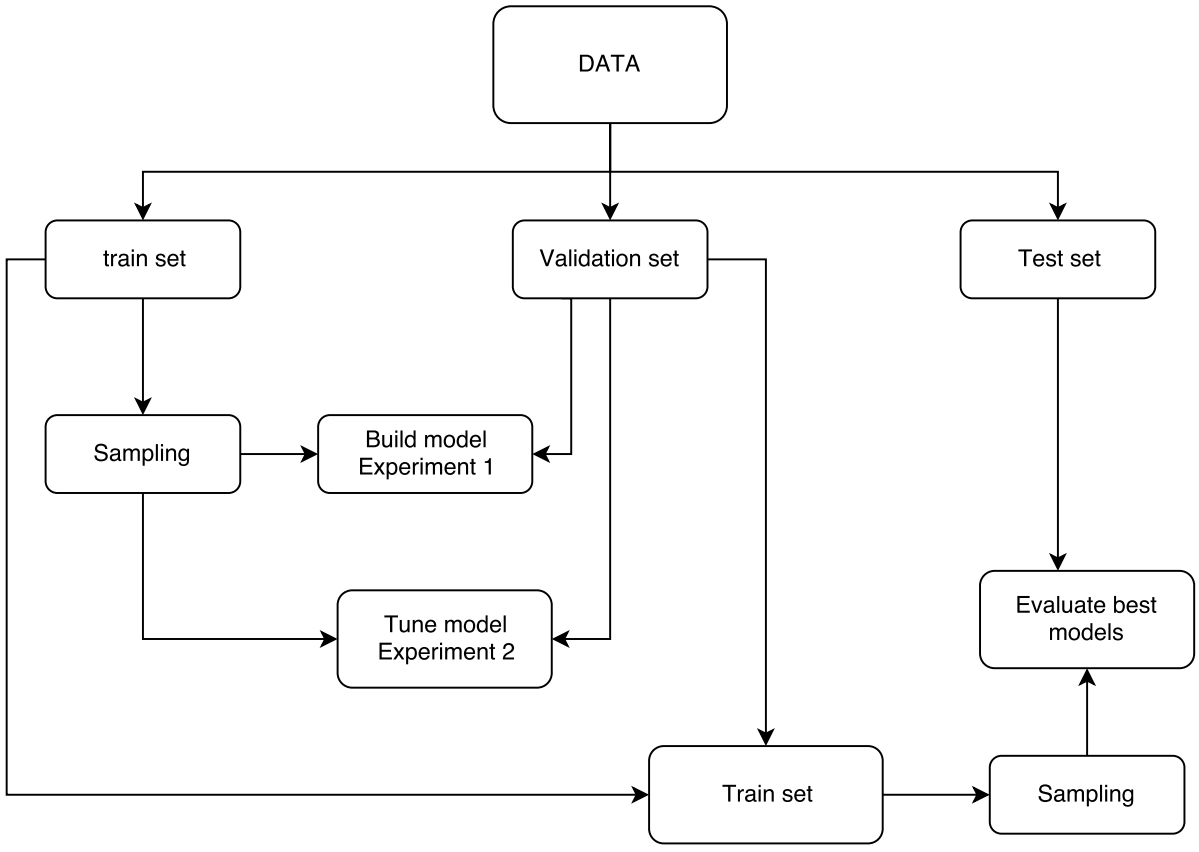


Figure 8: Model development experimental procedure

3.2.3 Identification of Factors Contributing to Student's Dropout

In order to identify and thorough understanding of factors that contribute significantly to the student dropout problem, we conducted 4 focus group discussions, 2 round table surveys, involving education stakeholders from our study area (10 teachers, 10 parents, 10 students, 5

education officers), the Director of Education Administration Division and 1 representative from the Twaweza. Topics of discussion: Student dropout problem (guided questions are summarized in Appendix 1). To ensure equal participation, each participant was given a chance to address the topic without interruption, followed by a discussion and a short debate. Data with similar responses were grouped, coded and analyzed using SPSS statistical software. Additionally, the study reviewed 122 articles in order to examine and compare the features that have already been reported by other researchers and features that were identified in our local context during Focus Group Discussion (FGD) and Round Table Survey (RTS). Permutation of the feature engineering was conducted to identify the contribution of each feature on the prediction performance by automatically selecting features that are most relevant to the dropout predictive model. This was accomplished by measuring permutation of the feature importance score (Pf_i) as defined in equation 1:

$$Pf_i = \mathrm{P_b} - \mathrm{P_s} \quad \ldots\ldots (1)$$

Where

(i) $\mathrm{P_b}$ is the base performance metric score

(ii) $\mathrm{P_s}$ is the performance metric score after shuffling

3.2.4 Data Preprocessing

The data from the three datasets were preprocessed prior to obtaining a final training set (Tables 3, 4 and 5). Data cleaning, normalization, transformation, feature extraction and selection were done. This was done as a precautionary measure to ensure datasets are well cleaned and accurate before model development. The data cleaning was done by removing information that could reveal the identity of individuals by the end user. Furthermore, the data samples with nominal features were converted to numerical values to conform with Scikit-learn. Dimensional reduction techniques which include Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Truncated Singular Value Decomposition (Truncated SVD) were employed to handle outliers (Fig. 9). The missing values were replaced using medians and zeros.

Table 3: Summary of the variables in Uwezo dataset

Variable description	Type of data
Main source of household income (Income)	Multinominal
Boy's Pupil Latrines Ratio (BPLR)	Numerical
School has girl's privacy room (SGR)	Binominal
Region	Nominal
District	Nominal
Village	Nominal
Student gender (Sex)	Binominal
Parent check child's exercise book once in a week (PCCB)	Binominal
Household meals per day (MLPD)	Multinominal
Student read any book with his/her parent in last week (SPB)	Binominal
Parent discuss child's progress with teacher last term (PTD)	Binominal
Student age (Age)	Numerical
Enumeration Area type (EAarea)	Multinominal
Household size (HHsize)	Numerical
Girl's Pupil Latrines Ratio (GPLR)	Numerical
Parent Teacher Meeting Ratio (PTMR)	Numerical
Pupil Classroom Ratio (PCR)	Numerical
Pupil Teacher Ratio (PTR)	Numerical
Dropout	Binominal

Table 4: Summary of the variables in India dataset

Variable description	Type of data
Continue drop	Binominal
Student id	Numerical
Gender	Binominal
Caste	Multinominal
Mathematics marks	Numerical
English marks	Numerical
Science marks	Numerical
Science teacher	Numerical
Languages teacher	Numerical
Guardian	Multinominal
Internet	Binominal
School id	Numerical
Total students	Numerical
Total toilets	Numerical
Establishment year	Numerical

Table 5: Summary of the variables in PORALG dataset

Variable description	Type of data
Region	Nominal
District	Nominal
Ward	Nominal
School name	Nominal
Dropout Male	Numerical
Dropout Female	Numerical
Pupil Teacher Ratio (PTR)	Numerical
Pupil Qualified Teacher Ratio (PQTR)	Numerical
Pupil Classroom Ratio (PCR)	Numerical
Boys Pupil Latrine Ratio (BPLR)	Numerical
Girls Pupil Latrine Ratio (GPLR)	Numerical

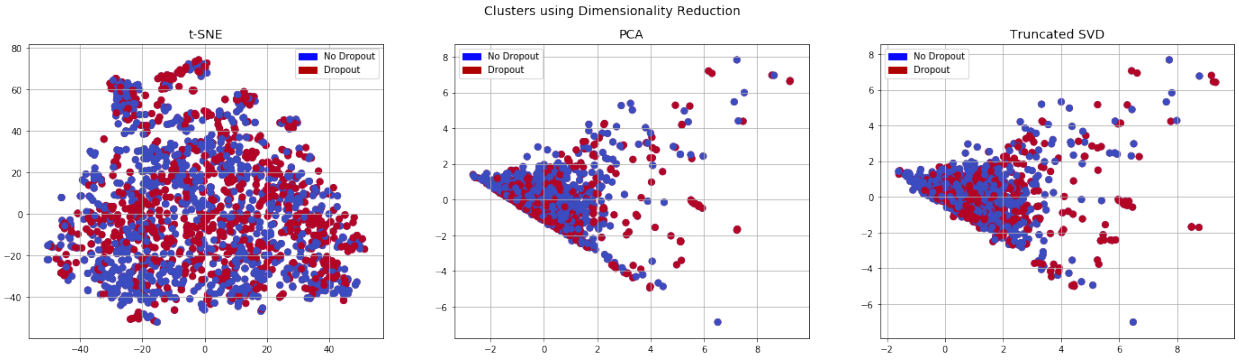


Figure 9: Dimensional reduction techniques

3.2.5 Data Balancing Techniques

To handle the issue of data imbalance in the datasets used in this study, three data balancing techniques were employed. Data balancing techniques before model development were selected due to their in-depth data cleaning, production of straight forward and satisfactory results when handling data imbalance, addressing the overfitting problem and reduction of running time and cost. (a) Under-sampling approach where Random Under Sampling (RUS) was used, (b) Over-sampling approach where Random Over Sampling (ROS) and Synthetic Minority Over-Sampling Technique (SMOTE) were used and (c) Hybrid approach where Synthetic Minority Over-Sampling Technique and Edited Nearest Neighbor (SMOTE ENN) and Synthetic Minority Over-Sampling Technique Tomek links (SMOTE TOMEK) were applied.

Random Under Sampling as the approach which tends to randomly select examples from the majority class for exclusion with no replacement until the outstanding number of examples were thoroughly together with that of the minority class was used. This approach was selected due to its ability to reduce the run time cost by decreasing the size of the data by eliminating some examples. Random Over Sampling was done by randomly balancing the distribution of data over the application of minority data duplication up to when the number of chosen examples plus the original examples of the minority class was roughly equal to that of the majority class. This approach was selected due to its ability of not eliminating important information in the data. Synthetic Minority Over-Sampling Technique was selected to form a new minority class examples by incorporating several minority classes examples. Furthermore, SMOTE TOMEK was selected to remove examples that form Tomek links from both classes and SMOTE ENN was selected to expel examples from both classes, therefore any example

that has been misclassified by its three nearest neighbors was removed from the training set. This technique was anticipated to give more in depth data cleaning because ENN have a tendency to eliminate more examples than Tomek links. The experimental procedure is summarized in Fig. 10.

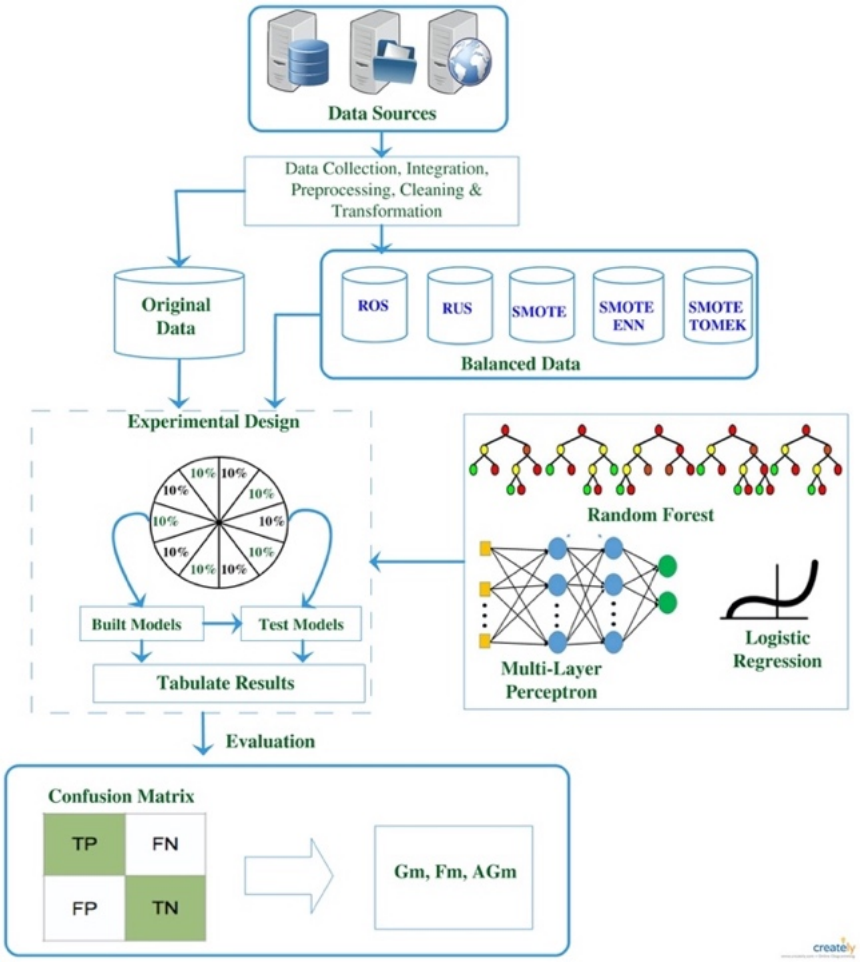


Figure 10: Imbalance techniques experimental procedure

3.2.6 Algorithms for Model Development

During model development, four supervised classification algorithms which are Linear Regression (LR), Multi-Layer Perceptron (MLP), Random Forest (RF) and K-Nearest Neighbor (KNN) were assessed on a set of supervised classification dataset in order to see which algorithms will perform better in machine learning model with consideration of the data imbalanced problem. Logistic Regression (LR) was selected to represent linear model and was used to model the probability of discrete outcome (either be binary or multinomial) and in this study it was binary outcome (dropout/not dropout). Additionally, RF represented ensemble model and was selected due to its ability to reduce the overfitting problem and handling high

dimensional data. The MLP on the other hand represented artificial neural network and was selected due to its ability to lower complexity and produce satisfactory results while K-NN represented instance model and was selected due to its simplicity and easy implementation.

Three evaluation metrics i.e. Geometric mean (G_m), F- measure (F_m) and Adjacent Geometric mean (AG_m) were used to examine the algorithms mentioned above. The selection of these metrics focused on imbalance domain and as a standard measure in class distribution. Geometric mean (G_m) was selected to measure the ability of the classifier to balance TP rates (sensitivity) and TN rates (specificity) as presented in equation 2. F- measure (F_m) was selected to measure the harmonic means of TP rates and precision as presented in equation 3 while AG_m was selected to measure the increase of TP rates without decreasing of TN rates as presented in equation 4. The best performed algorithms in all three metrics were then subjected to hyper-parameter optimization approach in order to improve the overall predictive power of the model. Tuned algorithms were then deployed voting technique to form an ensemble model. Furthermore, confusion matrix was evaluated to determine the best data balancing technique for the type of data used with consideration of the applied algorithms.

$$G_m = \sqrt{(\text{TPrate} \cdot \text{TNrate})} \qquad \text{.........} \ (2)$$

$$F_m = 2\text{PPV} \cdot \frac{\text{TPrate}}{\text{PPV} + \text{TPrate}} \qquad \text{.........} \ (3)$$

$$AG_m = \begin{cases} \frac{\text{GM}+\text{TNrate}\cdot(FP+TN)}{1+FP+TN} & \text{if TPrate} > 0, \\ 0 & \text{if TPrate} = 0 \end{cases} \qquad \text{.........} \ (4)$$

Where

(i) TN is true negative, TP is true positive, FN is false negative and FP is false positive.

(ii) $\text{TPrate} = \frac{\text{TP}}{\text{TP+FN}}$ the percentage of positive instances correctly classified.

(iii) $\text{TNrate} = \frac{\text{TN}}{\text{FP+TN}}$ the percentage of negative instances correctly classified.

(iv) $\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}$

3.2.7 Prototype Development

The prototype development followed software development approach (Fig. 11). This approach was created to receive feedback from users in order to refine the developed product. The aim was to develop a simplified version of the prototype and provide users with the evaluation and feedback. The prototype was then improved following feedback from the users. The improved prototype was given back to the users for further evaluation, and the cycle continued until the users were satisfied with the final prototype. The four phases of the prototype software development approach include:

(i) Requirements Gathering Process

Requirements gathering is the first step towards the design process. It involves identifying the users of the system, understanding and knowing their environment, what they do, and what they want to achieve via the system. There are two types of requirements, (a) functional requirements which refer to what the system will be able to achieve or perform, and (b) non-functional requirements which specify the attributes or constraint that the system must respect i.e. separate the requirements that focus on how good the software is from what the software is capable of doing.

(ii) Developing Alternative Designs

The second phase is design alternatives, which presents designs for the system as generated from the users' requirements. The first phase of the design process is called conceptual design, followed by the physical design. The conceptual design represents and validates the requirements gathered and is accomplished in collaboration between the designer and the users while the physical design focuses on the logical schema to constitute the actual physical structure of the database.

(iii) Building Interactive Version of the Designs

The third phase was to build prototype of the system that allows interaction with the users. This was built iteratively as reference was made to the requirements and the conceptual model.

(iv) Evaluating the Designs

Evaluation was carried out to make sure the final system is what was expected. This process was conducted in every iteration. The end users were involved in every step of the design process. This was done to make sure the users' context and requirements are incorporated in the system.

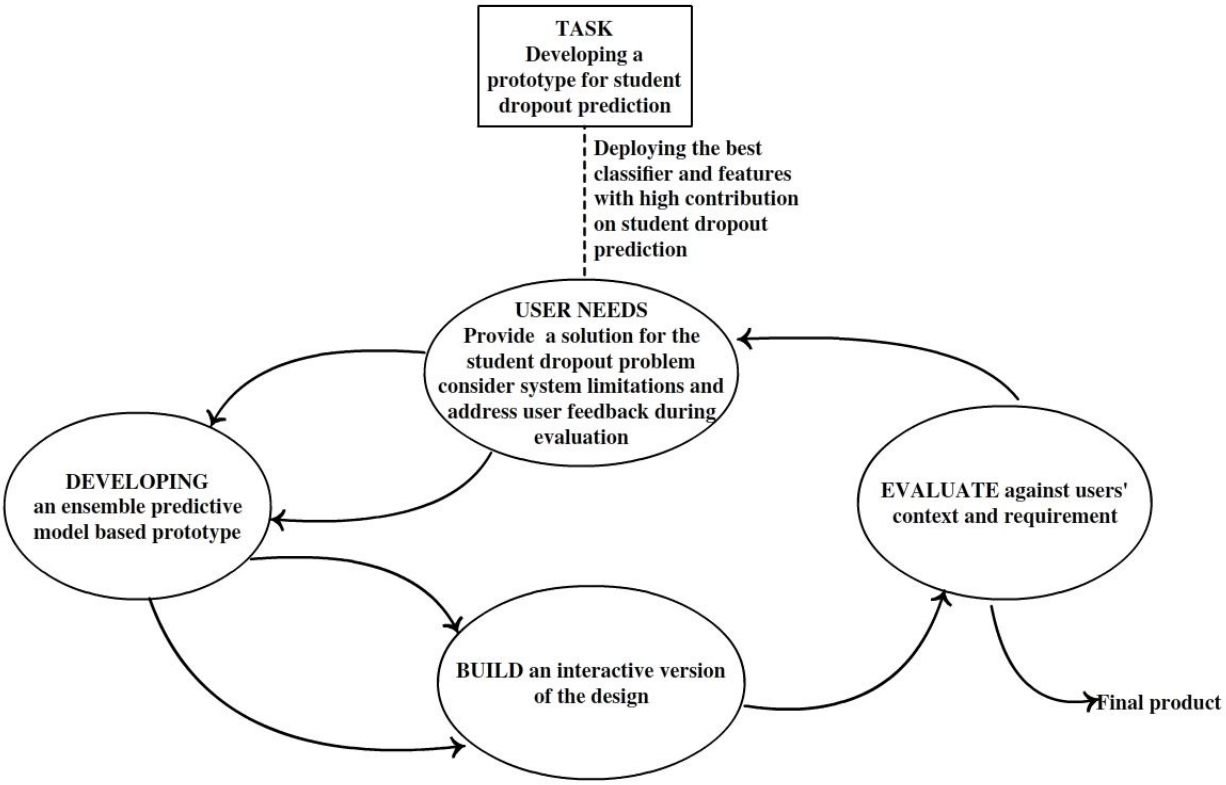


Figure 11: Prototype software development approach

3.2.8 Architectural System Design

The architectural system (Fig. 12) of developed prototype was designed based on the user's requirements. The prototype interface was linked to the server via the internet. The client side allowed input of the students' information, which was transferred through the system interface via the internet to the deployed model on the server. The model then predicts results for this new entry which was then being transferred via the internet to the prototype interface. The flask web server facilitated the record transfer to the server and the result from the server to the system interface while the heroku server was used as the hosting platform to support deployment of the developed system.

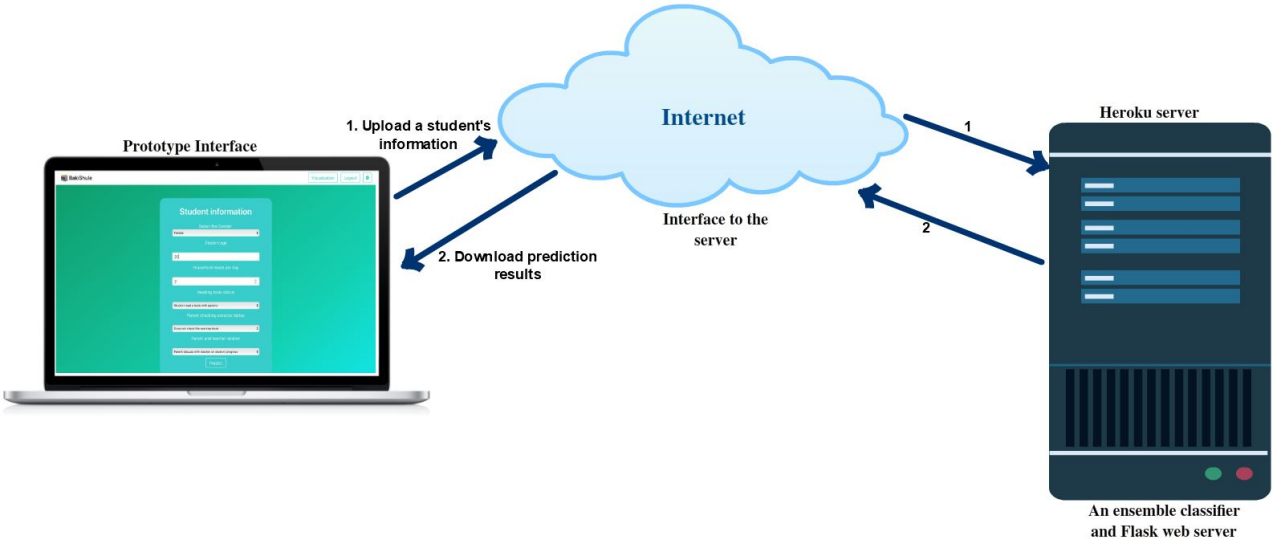


Figure 12: Architecture diagram of the system

3.2.9 Use Case Diagram

The use case diagram (Fig. 13) designed in this study describes users' interactions with the developed prototype. It gives a narrative description of the behavior of the system on its high level of abstraction. The system comprises of two main users; (a) administrators who were granted higher privileges than others, and (b) education stakeholders who were able to gain access to the system (after being authorized by an administrator) and input students' information, view predictive results, visualize school dropouts and upload new data entry for visualization (Table 6).

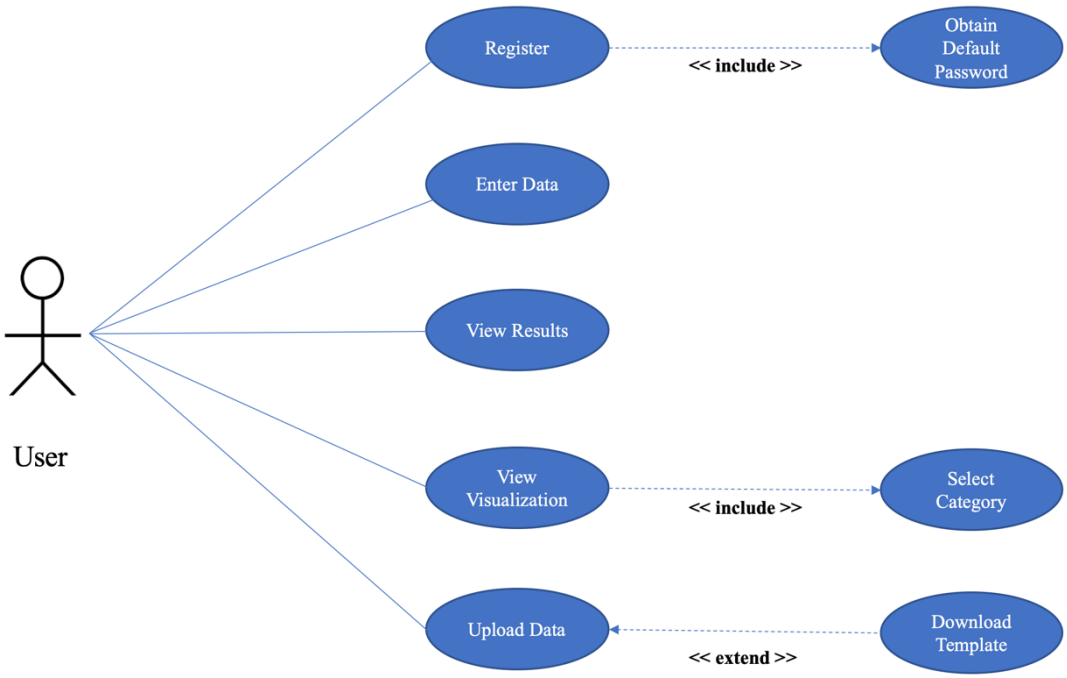


Figure 13: Use case diagram of the system

Table 6: User roles and system functionalities

User	Roles
Administrator	Input user
	Delete user
	Enter student input information
	View predictive results
	View visualizations
	Upload data for visualization
Education stakeholder	Enter student input information
	View predictive results
	View visualizations
	Upload data for visualization

3.2.10 Prototype Evaluation

(i) Technical Evaluation

A total of 10 systems development experts were invited to undertake technical evaluation on this system. Specifically, they were asked to evaluate BakiShule with consideration of accessibility, scalability, easy to use, consistency, navigation and feedback. Evaluators were required to rate each of the presented aspects into a scale as either Very High, High, Average, Low or Very Low by ticking one of the boxes and giving reasons for ratings below Average (Appendix 2). The number of evaluators was converted to percentage for easy graphical presentation.

(ii) End-user Evaluation

The end-user evaluation of the developed prototype was conducted by inviting similar education stakeholders (teachers, parents and education officers from Arusha, Mbeya, Nzega, Rufiji and Kisarawe districts) during the identification of factors that contribute to drop out to provide their evaluation feedback on the developed prototype. In this case, the following aspects were evaluated: ability of the system to predict whether a student will drop or not, ability of the system to visualize school with higher dropout risk, clarify of the predicted results and usefulness of the system.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

4.1.1 Factors Contributing to Student Dropout

The results from the FGD and RTS conducted with invited education stakeholders from the 5 districts showed that Student gender (Sex) was mentioned by the majority of the respondents (95%) to have a strong contribution to the student dropout problem followed by Number of Qualified teachers (89%) and Student Age (84%). On the other hand, only few respondents (8%) reported that the number of Health advisors and Crime rate have the strong contribution to the student dropout followed by Scores of peers (11%) (Fig. 14).

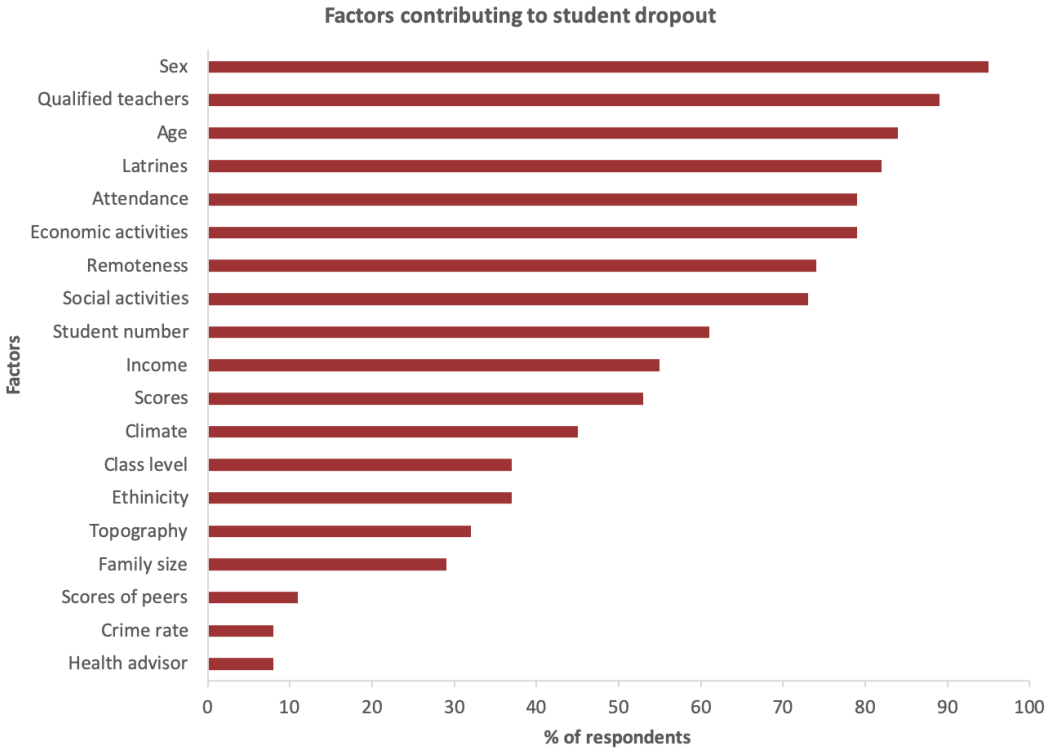
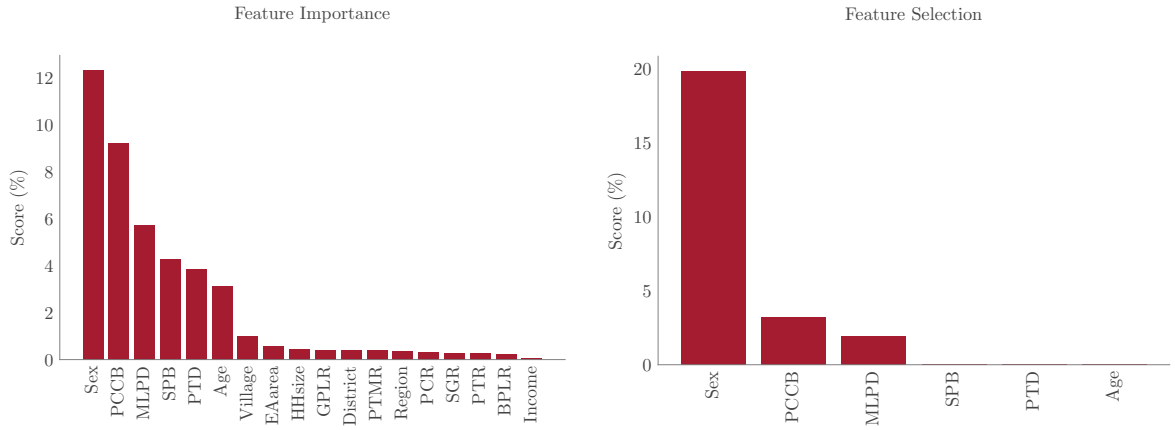


Figure 14: Factors contributing to student dropout identified during FGD and RTS

4.1.2 Feature Engineering

The results from the feature engineering experiment showed that Student gender (Sex) has strong contribution (12%) to student dropout followed by Parent who check his/her child's exercise book once in a week (PCCB) (9%), Household meals per day (MLPD) (6%), Student

who read any book with his/her parent in last week (SPB) (5%), Parent who discuss his/her child's progress with teacher last term (PTD) (4%) and Student age (Age) (3%) features (Fig. 8a). Other features such as Village, Household Size and so on had less contribution (< 1%) on the student dropout problem (Fig. 15a). The same experiment was repeated using the 6 well performed features out of the 18 features from the original dataset, and the results showed student gender (sex) have the strong contribution (20%) to the dropout prediction compared to the rest (Fig. 15b).

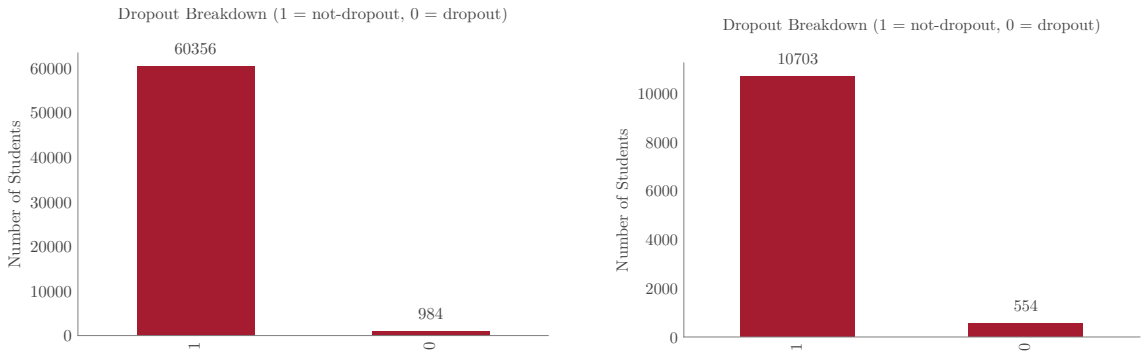


(a) (b)

Figure 15: Feature engineering experiment with all features (a), and with best performed features (b)

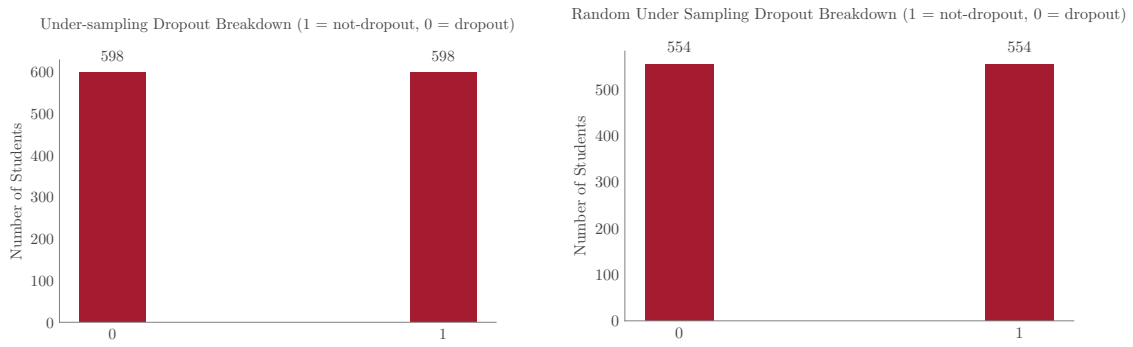
4.1.3 Data Sampling Balancing Techniques

The results from data sampling balancing techniques showed that, the dropout distribution for Uwezo dataset was 99.4% students retained and 1.6% students dropped out (Fig. 16a) while for India dataset, results showed that 95.08% retained and 4.92% dropped (Fig. 16b). Additionally, the results for each data balancing technique for both Uwezo (a) and India (b) are shown in Fig. 17 to 21. On the other hand, SMOTE ENN data balancing technique had very good solutions for accomplishing a greater performance using an original unsampled data, followed by SMOTE TOMEK and RUS on the Uwezo dataset (Table 7) while on the India dataset, SMOTE ENN data balancing technique performed better, followed by SMOTE TOMEK and ROS (Table 8). Furthermore, the three algorithms used in data balancing techniques were evaluated using confusion matrix and results showed that LR was the best algorithm to correctly classify the highest number of student dropout and misclassified the lowest followed by MLP and RF in both Uwezo (Table 9) and India (Table 10) datasets



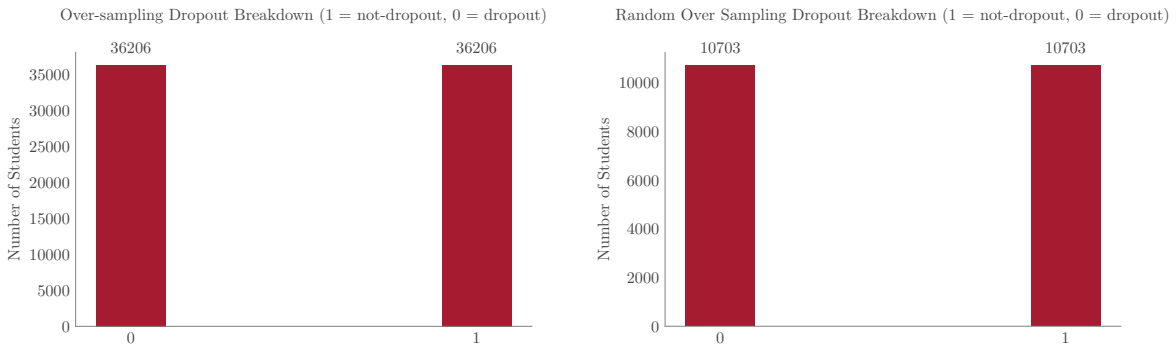
(a) (b)

Figure 16: Dropout distribution using Uwezo dataset (a) and India dataset (b)



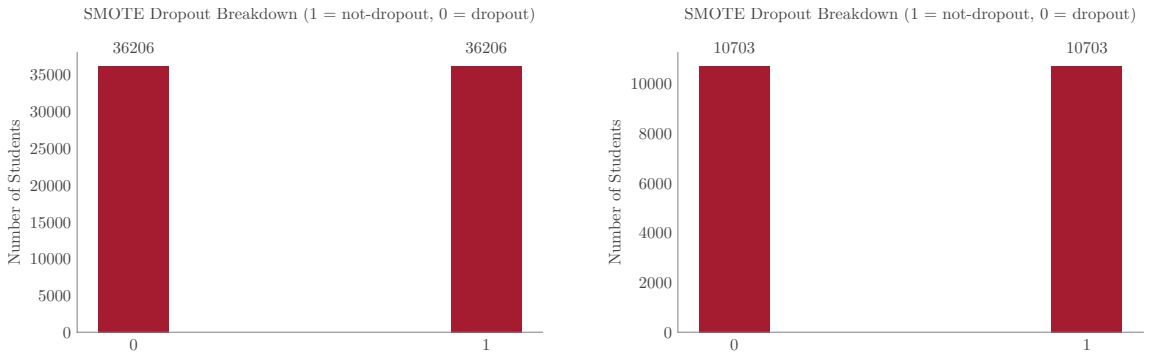
(a) (b)

Figure 17: Dropout distribution using RUS for Uwezo dataset (a) and India dataset (b)



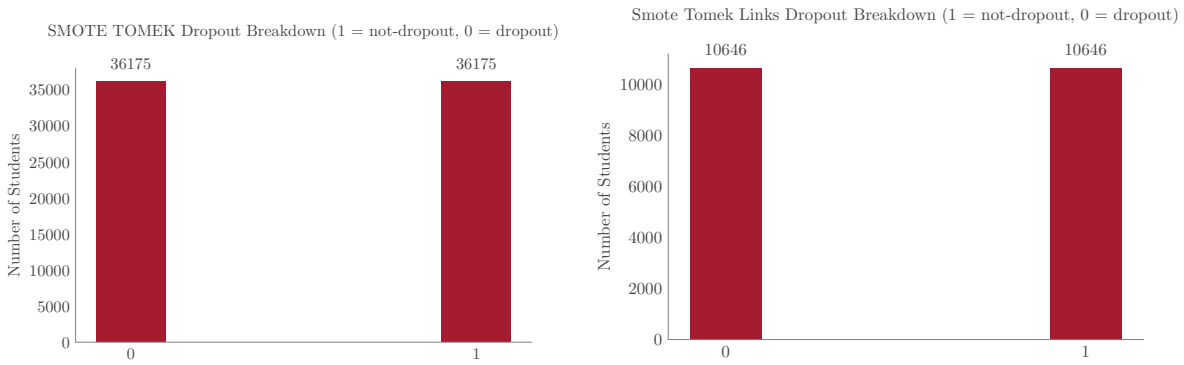
(a) (b)

Figure 18: Dropout distribution using ROS for Uwezo dataset (a) and India dataset (b)



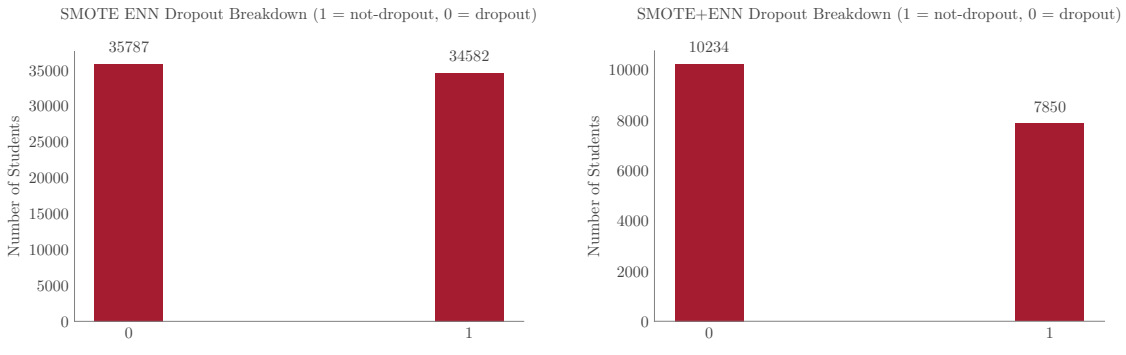
(a) (b)

Figure 19: Dropout distribution using SMOTE for Uwezo (a) and India (b) datasets



(a) (b)

Figure 20: Dropout distribution using SMOTE TOMEK for Uwezo dataset (a) and India dataset (b)



(a) (b)

Figure 21: Dropout distribution using SMOTE ENN for Uwezo dataset (a) and India dataset (b)

Table 7: Summary of experimental results for Uwezo dataset

Preprocessing	Models	G_m	F_m	AG_m
None	LR	0.000	0.000	0.000
	MLP	0.011	0.002	0.012
	RF	0.004	0.000	0.004
Random over sampling	LR	0.536	0.547	1.010
	MLP	0.499	0.438	0.920
	RF	0.293	0.270	0.449
Random under sampling	LR	0.548	0.546	1.042
	MLP	0.512	0.332	1.031
	RF	0.624	0.561	1.192
SMOTE	LR	0.551	0.556	1.034
	MLP	0.525	0.475	0.967
	RF	0.661	0.645	1.138
SMOTE + ENN	LR	0.562	0.572	1.079
	MLP	0.577	0.491	1.104
	RF	0.676	0.666	1.176
SMOTE + Tomek	LR	0.550	0.556	1.032
	MLP	0.546	0.508	1.015
	RF	0.663	0.646	1.140

Table 8: Summary of experimental results for India dataset

Preprocessing	Models	G_m	F_m	AG_m
None	LR	0.000	0.000	0.000
	MLP	0.000	0.000	0.000
	RF	0.031	0.002	0.031
Random over sampling	LR	0.616	0.592	1.137
	MLP	0.525	0.450	0.957
	RF	0.707	0.667	1.207
Random under sampling	LR	0.582	0.570	1.085
	MLP	0.515	0.139	0.925
	RF	0.711	0.667	1.210
SMOTE	LR	0.648	0.603	1.190
	MLP	0.555	0.410	1.032
	RF	0.707	0.667	1.207
SMOTE + ENN	LR	0.722	0.638	1.343
	MLP	0.791	0.438	1.531
	RF	0.738	0.706	1.283
SMOTE + Tomek	LR	0.655	0.605	1.201
	MLP	0.735	0.441	1.390
	RF	0.707	0.667	1.207

Table 9: Comparison of algorithms for Uwezo dataset

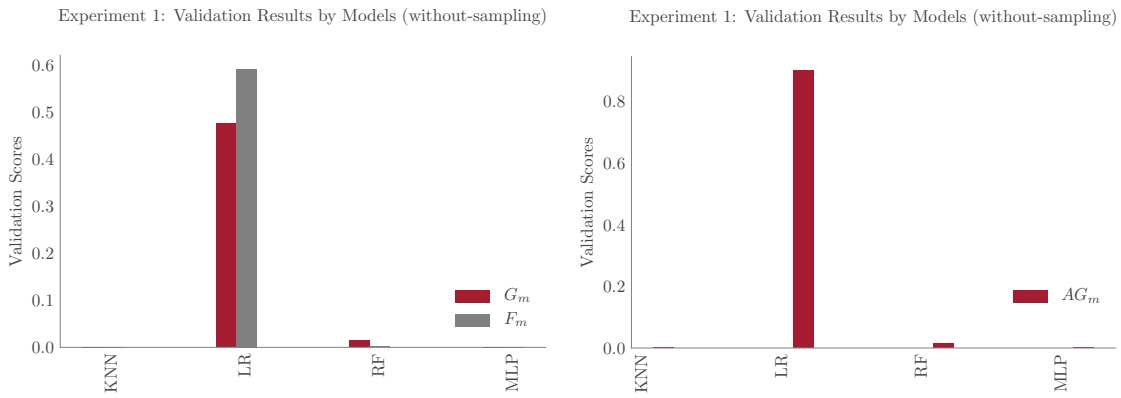
Item	LR	MLP	RF
Students correctly classified	68 069	66 972	65 089
Students incorrectly classified	2300	3397	5280
Correctly classified dropout students	57 348	56 540	54 910
Incorrectly classified dropout students	1090	1111	1278
Correctly classified non-dropout students	10 721	10 432	10 179
Incorrectly classified non-dropout students	1210	2286	4002

Table 10: Comparison of algorithms for India dataset

Item	LR	MLP	RF
Students correctly classified	14 404	13 673	12 761
Students incorrectly classified	3680	4411	5323
correctly classified dropout students	13 430	12 781	11 972
incorrectly classified dropout students	698	791	981
correctly classified non-dropout students	974	892	789
incorrectly classified non-dropout students	2982	3620	4342

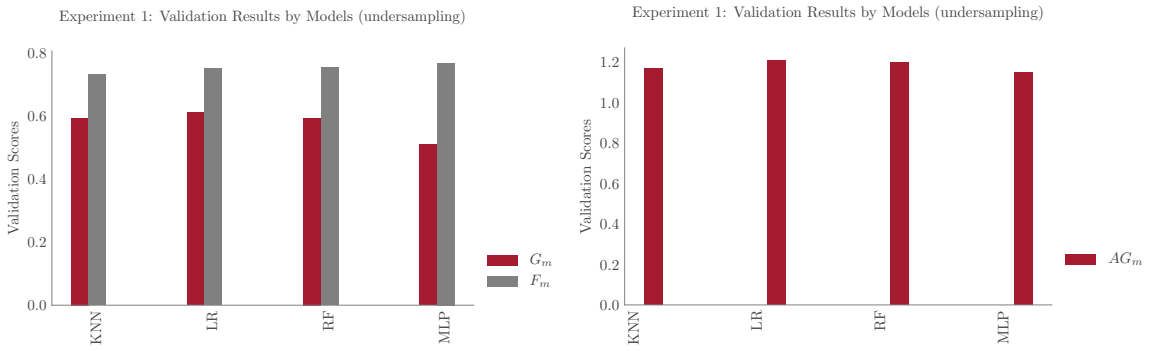
4.1.4 Model Development

During model development, four supervised learning algorithms (LR, MLP, RF and KNN) were evaluated using G_m, F_m and AG_m metrics to assess model performance in no sampling, under sampling and hybrid sampling techniques. The results showed that with no sampling technique, LR algorithm showed better performance than other algorithms (Fig. 22). In the case of under sampling, results indicated that all algorithms (LR, MLP, RF and KNN) had similar performance (Fig. 23) while in hybrid sampling LR and MLP algorithms showed better performance compared to RF and KNN (Fig. 24).



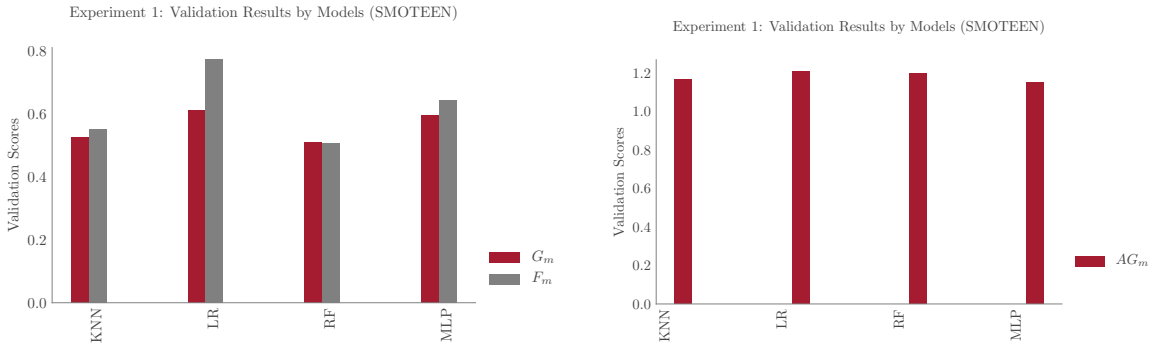
(a) (b)

Figure 22: No sampling validation results for G_m and F_m (a) and AG_m (b)



(a) (b)

Figure 23: Under sampling validation results for G_m and F_m (a) and AG_m (b)



(a) (b)

Figure 24: Hybrid sampling validation results for G_m and F_m (a) and AG_m (b)

4.1.5 Model Predictive Performance

In this experiment, two best performed algorithms (LR and MLP) were selected. The selected algorithms were then tuned using hyper-parameter optimization technique to further improve their predictive performance with consideration of the best parameters (Table 11). Furthermore, staking technique was employed to softly combine the tuned algorithms (LR2 and MLP2) to form an ensemble model (ENB). Results showed that, the performance of the tuned algorithms was significantly improved in both validation (LR2: AG_m= 1.372 and F_m= 0.894) and test (LR2: G_m= 0.783) scores compared to its baseline settings (Table 12), and an ensemble model (ENB) showed considerably better performance in both validation (G_m = 0.735) and test (AG_m= 1.335 and F_m= 0.847) scores (Table 12).

Table 11: Parameters considered during model tuning

Algorithm	Parameter
LR	fit_intercept: True, tol:1, C: 0.001, Penalty: "l1"
MLP	solver: "adam", learning_rate_int: 0.001, shuffle: True, hidden layer_size: 10, alpha: 1, early_stopping: True

Table 12: Hyper-parameters optimization

		LR	LR2	MLP	MLP2	ENB
Validation score	G_m	0.724	0.726	0.613	0.711	0.735
	AG_m	1.261	1.372	1.211	1.324	1.370
	F_m	0.841	0.894	0.723	0.827	0.891
Test score	G_m	0.721	0.783	0.621	0.706	0.779
	AG_m	1.320	1.332	1.278	1.281	1.335
	F_m	0.823	0.831	0.726	0.732	0.847

4.1.6 Users' Requirements

The results from the FGD and RTS conducted with the end users showed that student information input (one student at a time) was required by the majority of the respondents (97%) rather than uploading files. Additionally, 95% of the respondents required a prototype to be accessed at any time using either mobile phone or computer. On the other hand, 61% of the respondents saw a need for the developed prototype to be scaled up later when needed while 71% of the respondents required the system to allow users to navigate easily from one interface to the next. (Fig. 25).

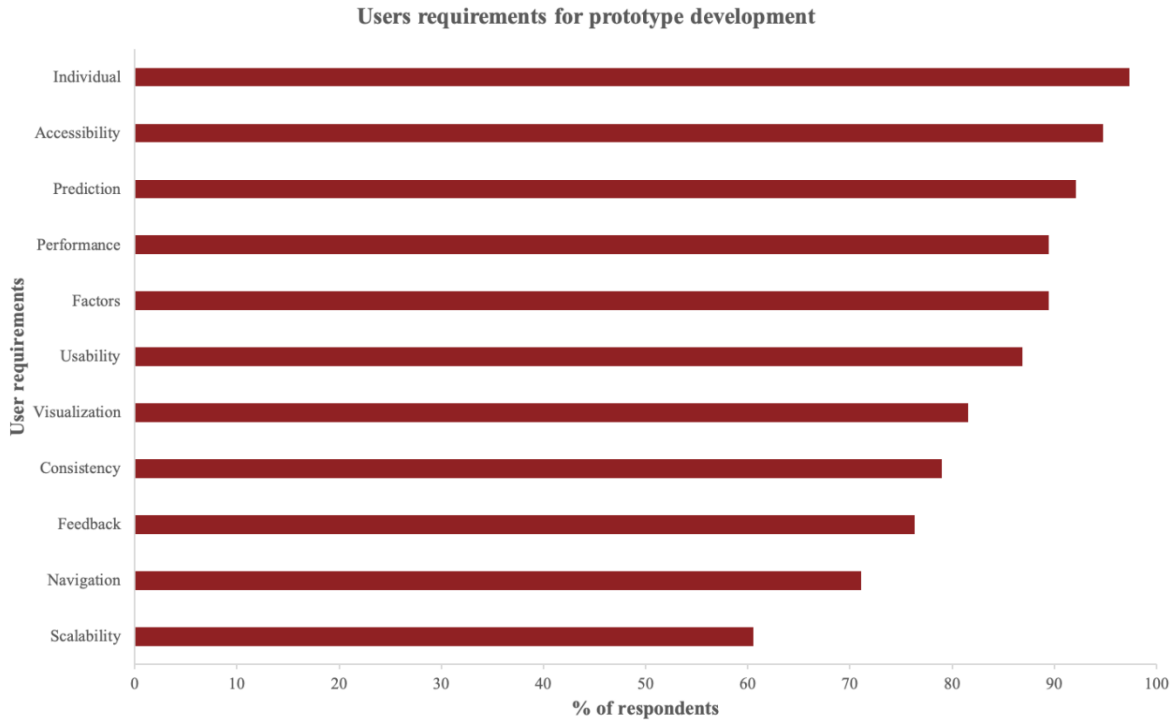


Figure 25: Users' requirements for prototype development

4.1.7 Prototype Development

The developed prototype (BakiShule) consist of login interface (Fig. 26) where a user is required to sign up and log in, input information interface (Fig. 27) for filling up student information (one student at a time) based on the 6 selected factors (Fig. 15b) which differs per individual student, predictive result interface (Fig. 28) which gives out the summary of the students information and the predictive result in the form of status (Dropout/Not Dropout). Additionally, visualization module was integrated into the developed prototype in order to visualize dropout rate in schools based on the region or district for education stakeholders to

easily understood and compare the dropout rates. Users were able to visualize total dropout (Fig. 29), dropout by gender (Fig. 30) and dropout by enrollment (Fig. 31) based on the selected region. Additionally, users were having a chance to upload a new dataset for visualization (Fig. 32) as guided by the provided template (Fig. 33).

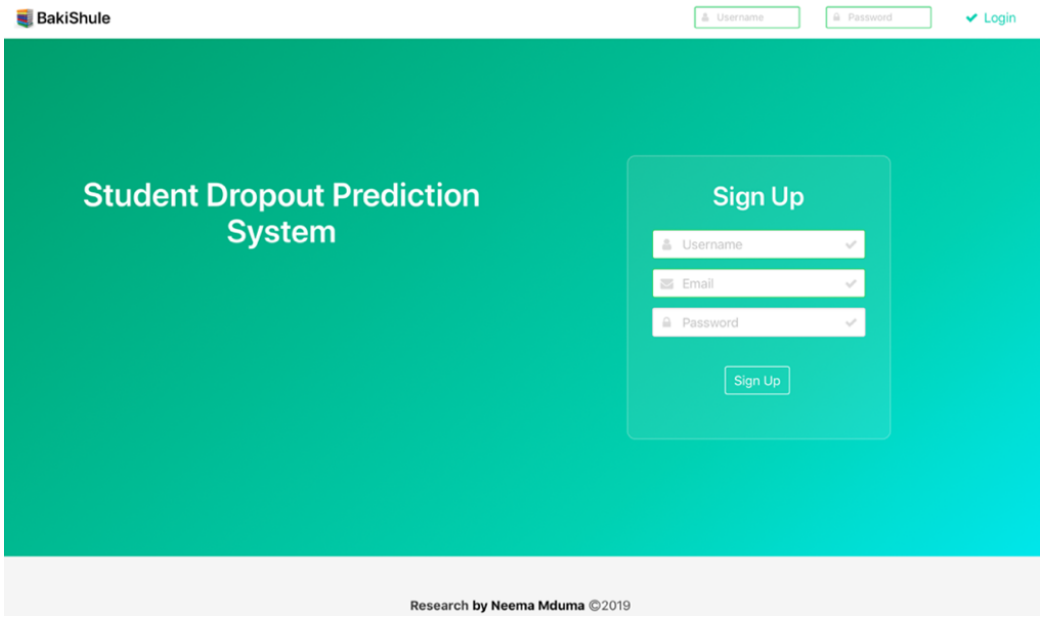


Figure 26: Login interface

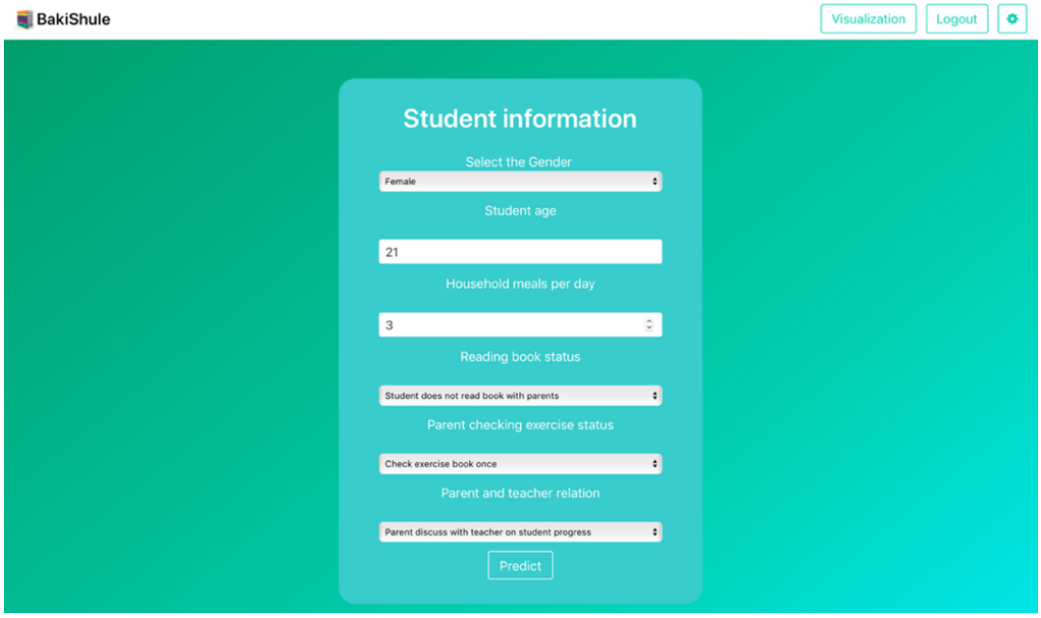


Figure 27: Input information interface

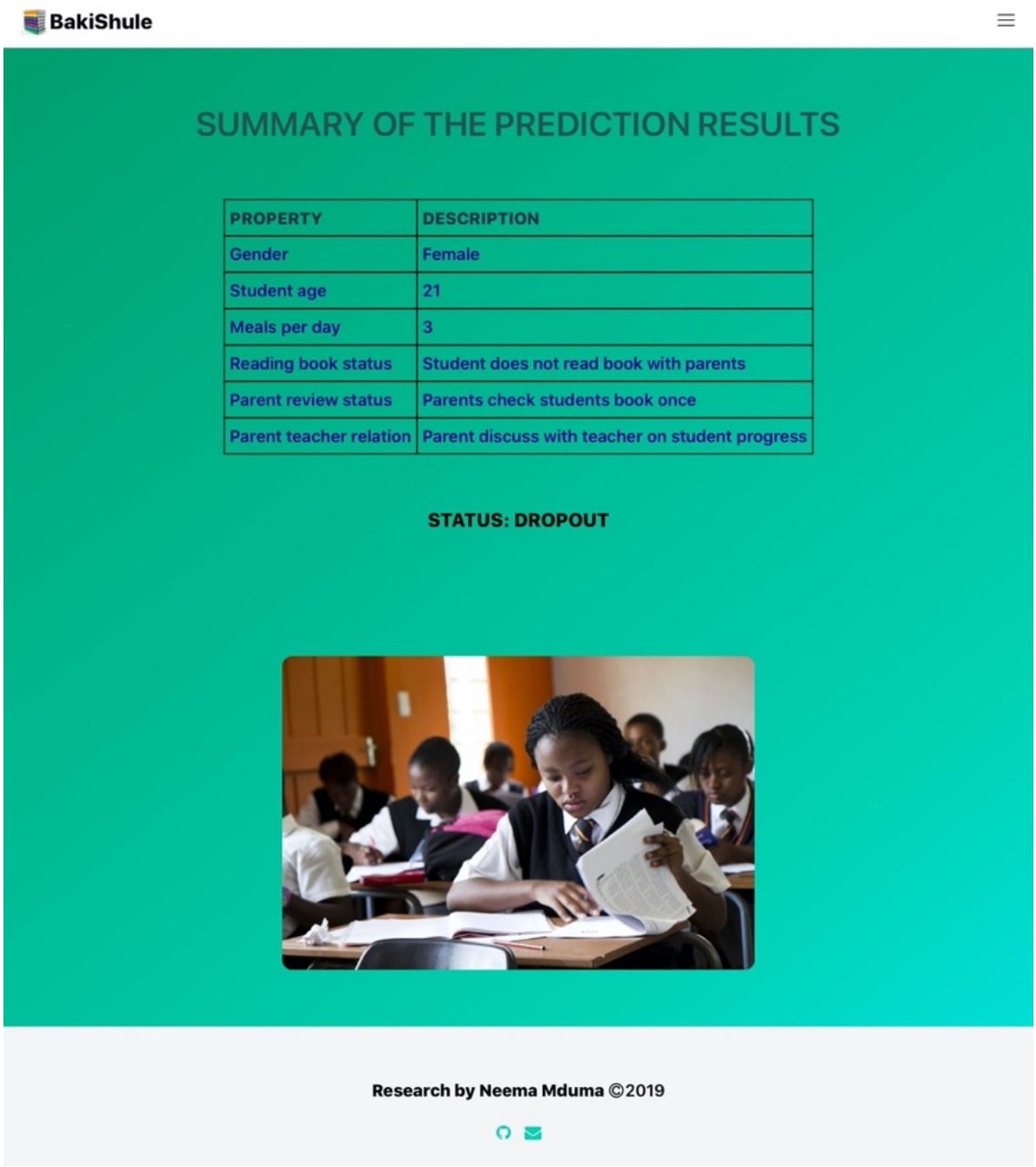


Figure 28: Predictive result interface

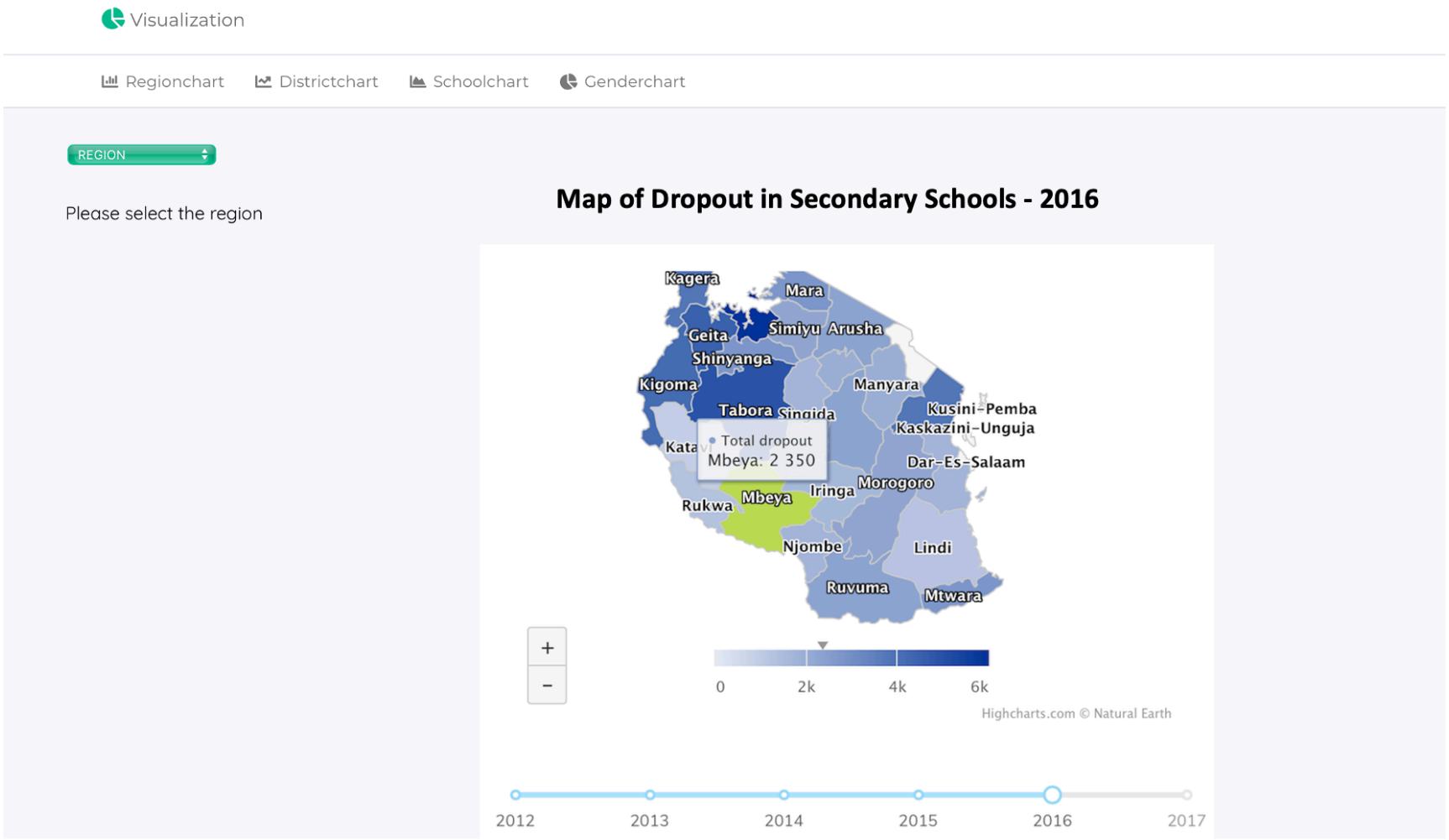


Figure 29: Visualization of total dropout of a selected region

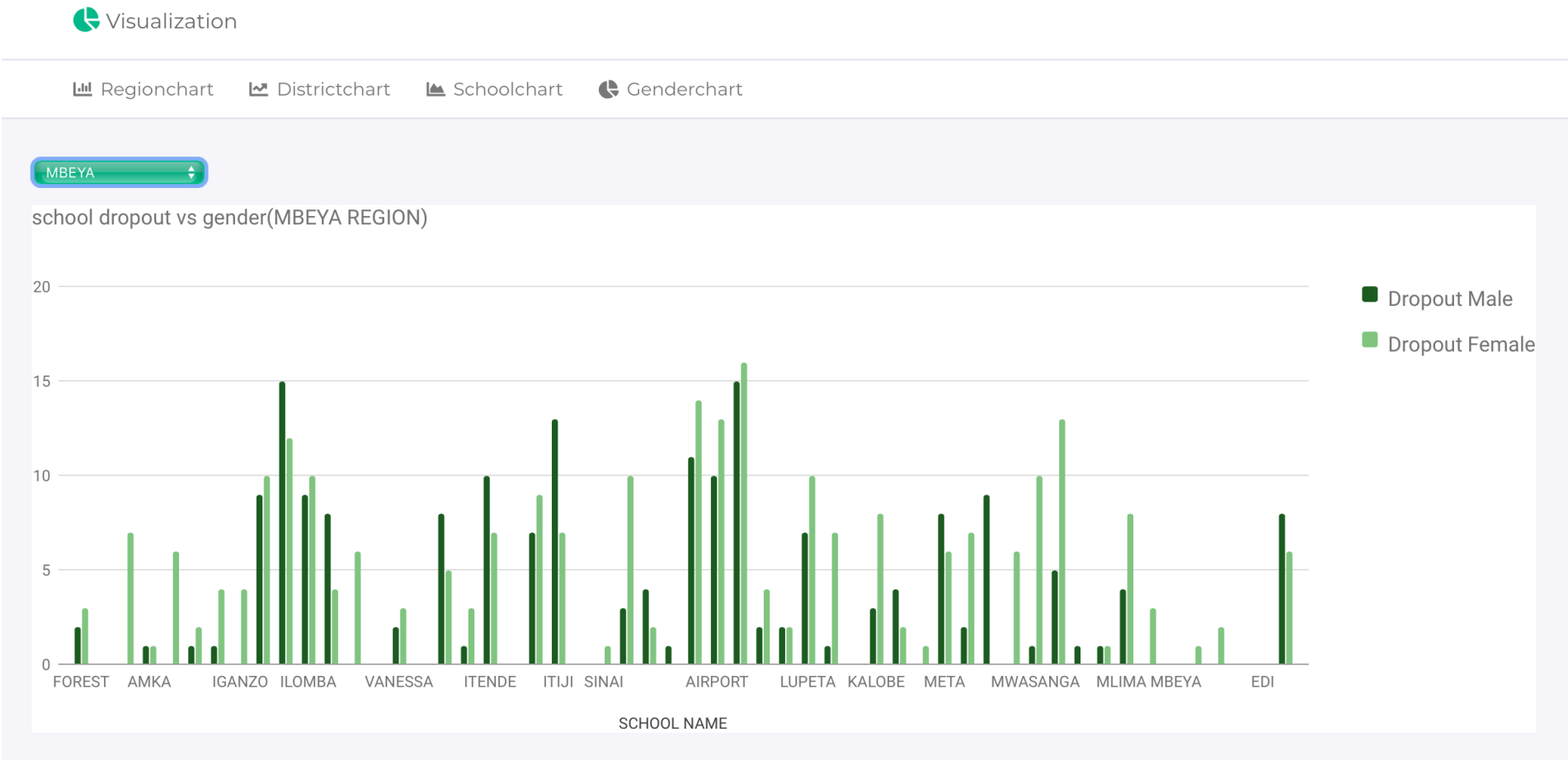


Figure 30: Visualization of dropout by gender of a selected region

Figure 31: Visualization of dropout against enrollment of a selected region

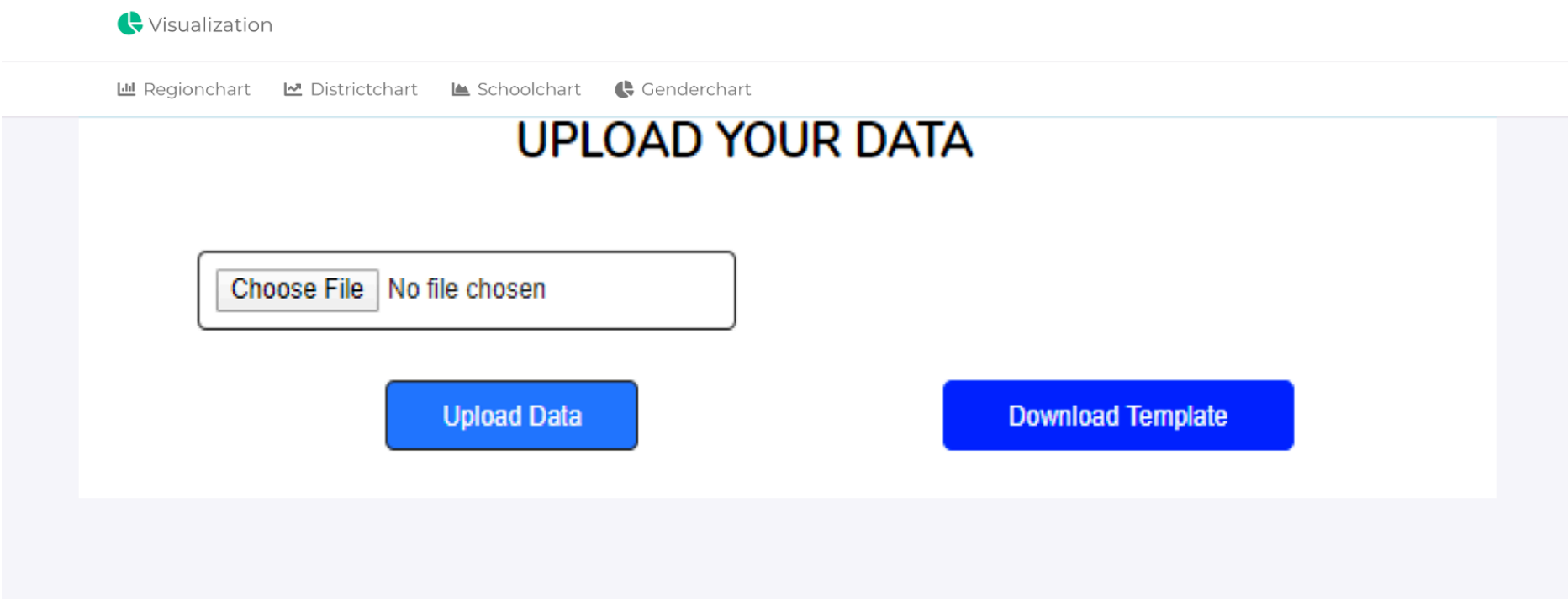


Figure 32: The upload data section for visualization

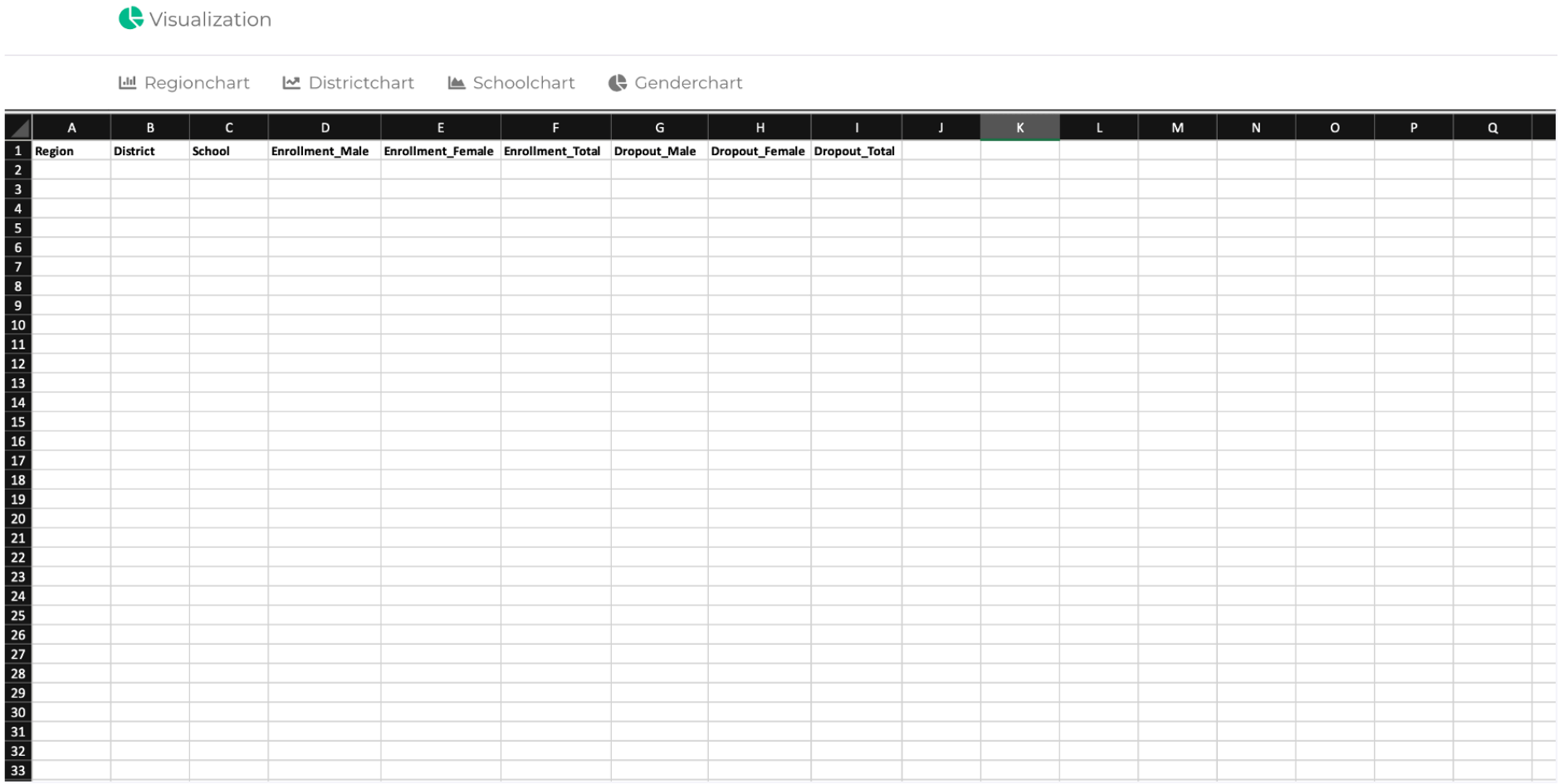


Figure 33: The data entry template for visualization

4.1.8 Prototype Evaluation

The evaluation of the developed prototype was grouped into technical and end-user. Technical and end-user evaluators were required to rate each of the presented aspects into a scale of either Very High, High, Average, Low or Very Low. Aspects to be evaluated by technical experts were Accessibility, Scalability, Usability, Consistency, Navigation and Feedback while end-users assessed ability of the system to predict student dropout, the ability of the system to visualize school dropout, clarity of the predictive results and usefulness of the system. Results showed that 90% of the technical experts rated the system very high in scalability, 80% rated system navigation capability very high as well (Fig. 34). On the other hand, 10% of the technical experts rated the system performance low in feedback category. Results from end-user evaluations showed that, 90% of evaluators were satisfied with the systems' ability to visualize school dropout and 80% ranked the system performance very high on its ability to predict student's dropout (Fig. 35). Despite the good evaluation from the end users, about 10% were not satisfied with the clarity of the predicted results.

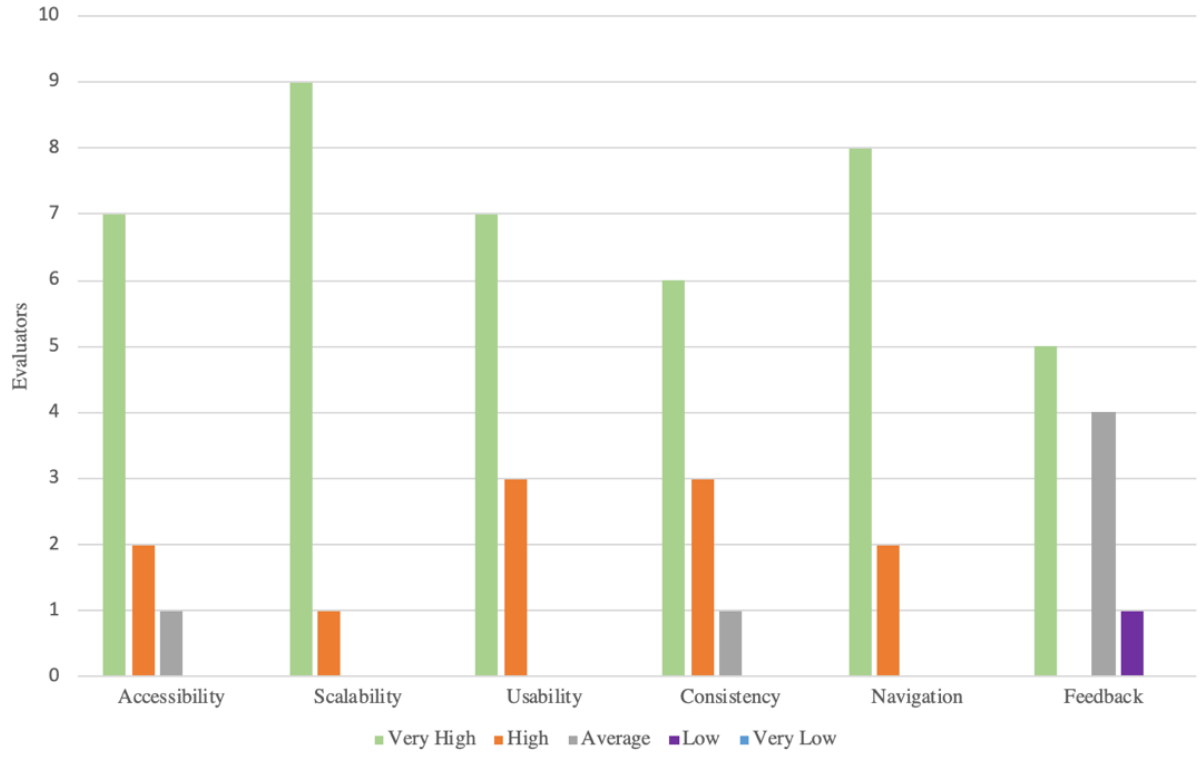


Figure 34: Technical expert's evaluation results

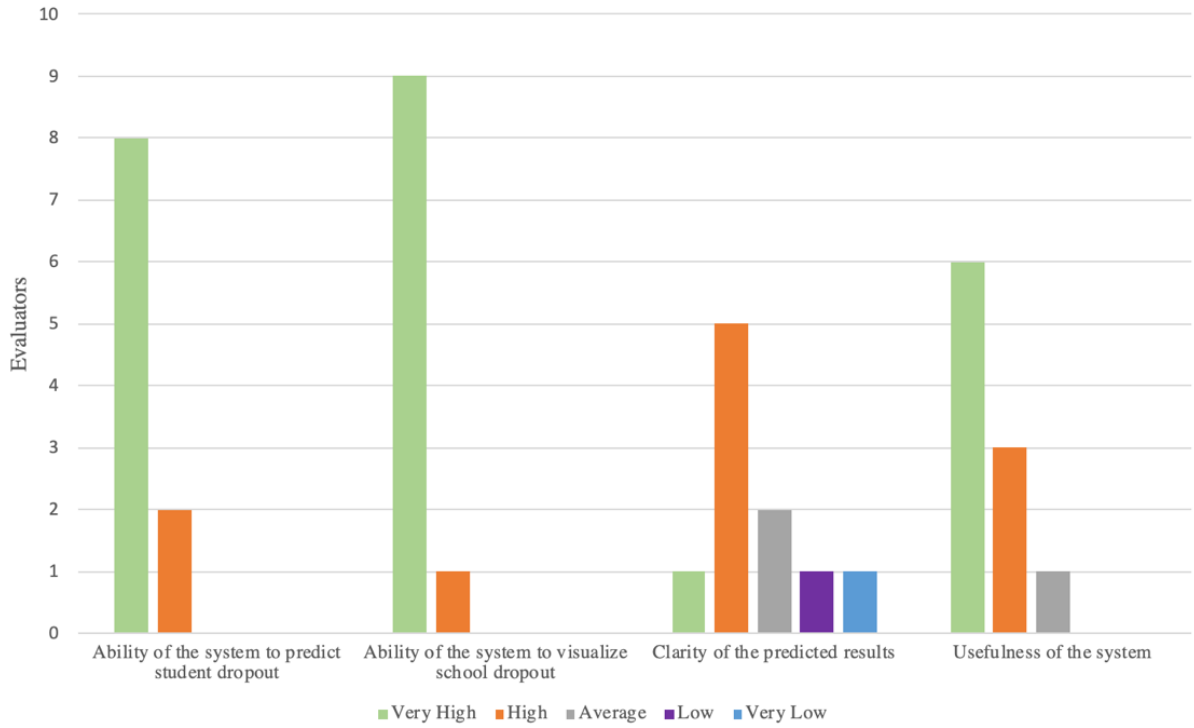


Figure 35: End-users evaluation results

4.2 Discussion

The majority of respondents in this study mentioned students' gender as a main contributing factor to the dropout problem, where girl's dropout rate in secondary schools was higher compared to boys. Similar results were previously reported by the BEST (2015) that the dropout rate in the country is pronounced to be 30% girls and 15% boys, which further confirms that girl's dropout rate is higher i.e. twice as much in this case as compared to boys. Additionally, most studies reported student gender as the main contributing factor of student dropout particularly in developing countries, and only a handful of studies found this to be the problem in developed world (Isphording & Qendrai, 2019; Kim *et al.*, 2018). The reasons for the higher dropout rate for girls was linked mostly to cultural factors such as early marriage and early pregnancies. The report from UNESCO (2017) revealed that early pregnancy and marriage accounts for 47% of girls' dropout in Sub-Saharan Africa, and further suggested that, if serious measures are not taken to rescue this situation, then the dropout rate can increase to 70% by the year 2030. This can have serious impact on attaining sustainable development goal 4 which stresses the need for inclusive and equitable quality education to all children of school age particularly girls from the marginalized communities.

On the other hand, crime rates in a locality and lack of health advisors in schools were mentioned as the least contributing factors to dropout problem by the majority of the respondents in our study. These results, however contradict with the findings reported by Injendi (2014), Bäckman (2017) and Bell *et al.* (2017) were high crime rates were associated with student dropout. Lack of school health advisors was also reported by Funja (2018) as one of the contributing factor to the student dropout problem particularly to girls of adolescent age. The two factors were thought to be the least contributing to the student dropout problem due to communities' perceptions and limited knowledge on how they directly contribute to the student dropout problem.

The feature engineering experiment conducted in this study showed that student gender (Sex) had the strong contribution to student dropout followed by parent who check his/her child's exercise book once in a week (PCCB) and household meals per day (MLPD), while features such as household size (HHS) and income were found to contribute less to the dropout problem. The majority of the features identified by the conducted experiment such as PCCB and MLPD were neither mentioned by the respondents nor found in the social science studies reviewed. Most social science studies and the communities identified factors using prior knowledge and in most cases the factors mentioned are the one with direct and obvious link to the dropout problem and ignore the one with an indirect link to the problem. Nevertheless, these traditional approaches of identifying contributing factors failed to address the problem over the years, and dropout seems to be increasing particularly in developing countries (Ananga, 2011; Lewin, 2009; UNESCO, 2017). On the other hand, machine learning model developed in this study was able to identifying factors that were ignored or believed to have no or little contribution to the dropout problem by the community and other researcher. This was made possible by the models' ability to mine non-linear information from the features or variables and their association with student dropout. These results are similar to what Khandani *et al.* (2012) and Gambacorta *et al.* (2019) reported in the USA and China on consumers-credit risk models. These findings suggest that ML models could help in the selection of non-linear factors that may be very essential on addressing the student dropout problem particularly in developing countries.

This study used 3 datasets to train the machine learning model on predicting student dropout, compare data balancing techniques and to visualize school dropout. The datasets used were highly imbalanced due to the fact that students who retain studies are always many compared

with students who dropout, thus making data balancing very important in this study as the main focus was on the minority class in this case students who dropout of school. The SMOTE ENN data balancing technique showed very good solutions for accomplishing a greater performance due to its ability to give more in depth data cleaning. Similar results were noted by Batista *et al.* (2004) when assessed several methods for balancing machine learning training data. Furthermore, Farquad and Bose (2012) stressed on the techniques and importance of handling data imbalance when developing training sets from machine learning model, and Ramentol *et al.* (2012) emphasized on the good performance of hybrid data balancing techniques such as SMOTE-RSB, SMOTE TOMEK and SMOTE ENN when dealing with highly imbalanced data like in our case of student dropout.

On the contrary, the RUS technique showed the least performance during data balancing evaluation experiment conducted in this study. This could be due to its nature of losing some potential information that might impact the learning process. Similar results were noted by Yen and Lee (2009) and Wang and Yao (2013) when assessed several approaches for handling imbalance datasets. However, this approach was reported to improve the predictive performance in other studies as compared to not using any data sampling techniques (Burez & Van den Poel, 2009; Prusa *et al.*, 2015). Most datasets in real world are not balanced i.e. there is majority and minority class, and if data balancing is ignored when training the machine learning model, it may lead to biasness towards one class and the model will learn more about majority class and learn less or ignore the minority class hence handling unbalanced data is very important when develop machine learning model.

Additionally, machine learning algorithms used in this study were evaluated using G_m, F_m and AG_m as metrics generated from confusion matrix. The basis for the selection of these evaluation metrics was due to their ability to perform well in the imbalance domain as compared to other metrics. Similar metrics were used by Aulck *et al.* (2017), Batuwita and Palade (2012), Kim *et al.* (2015), Kuncheva *et al.* (2019), Mgala (2016) and Rovira *et al.* (2017) in evaluating performance of the developed algorithms in order to take into account the class distribution. Furthermore, accuracy has been reported as the commonly used metric for measuring the degree of correctness in machine learning models (Ameri *et al.*, 2016; Lakkaraju *et al.*, 2015). However, its limitations in the imbalanced domain makes it unsuitable for evaluating model with imbalanced data (Lin & Chen, 2012; Longadge *et al.*, 2013; López *et al.*, 2013).

Moreover, our study noticed that the Logistic Regression and Multi-Layer Perceptron were the best performed algorithms to correctly classify the highest number of student dropout and misclassified the lowest. This could be due to the ability of Logistic Regression to model the probability of discrete outcomes (binary for the case of this study) and the power of Multi-Layer Perceptron to produce satisfactory results for non-linear relationships. Similar results were reported by Hakim (2019) and Mgala and Mbogho (2015b) when determining the accuracy of their predictive models for early prediction of brain strokes and students dropout respectively. Both studies reported LR as the best performed classification algorithm as compared to others. These results, however, contradicts to what was reported by Amin and Ali (2017) in their study of evaluating performance of supervised machine learning algorithms in healthcare, where K-Nearest Neighbor and Random Forest were reported to outperform other algorithms such as Logistic Regression and Naïve Bayes. To increase predictive performance, this study combined the two best performed algorithms (Logistic Regression and Multi-Layer Perceptron) to form an ensemble classifier. A similar technique was also applied by Zubair and Zahid (2019) to achieve better performance of their model for predicting chronic kidney disease by combining multiple learning algorithms to form an ensemble classifier.

The prototype developed in this study took into consideration the end users' requirements, and was deployed into a user interface prototype for end users to easily interact with. The majority of the respondents wanted a system that will allow student input information one at a time rather than uploading files with a list of students. This could be due to the number of students dropping out is always few compared to continuing students, hence making it easy to correctly filling in their information case by case in the class or school. Additionally, parents who are among the end users were only interested with their children's school progress, thus recommended a system that will allow them to fill in the information of their children one at a time for better follow up and prediction. Similar results were reported by Mgala (2016) when developing a mobile application to predict students' academic performance in Kenya specifically based on users' requirements. Additionally, Matto and Mwangoka (2018) and Maginga *et al.* (2018) stressed on the importance of considering the users' needs when developing the system. Furthermore, the developed prototype was again evaluated by technical experts and end users on its overall performance and ability to predict student dropout and gauge if it meets their requirements and expectations. Technical and end-user evaluation are important and provides room for improvement based on the feedback (Neumann, 2004; Warfel, 2009).

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study focused on developing a data driven approach for predicting student dropout in secondary schools in Tanzania. Permutation of the feature engineering experiment was conducted in a set of 18 features to identify features with the strong contribution to the dropout prediction. Five data balancing techniques were compared using three popular classification techniques (LR, MLP and RF) alongside with real world featured datasets from Tanzania and India. Furthermore, four supervised classification algorithms (LR, MLP, RF and KNN) were empirically assessed on a set of supervised classification dataset in order to provide a contemporary set of recommendations to researchers who wish to apply machine learning algorithms to their data with consideration of the data imbalanced problem.

The developed model was deployed into a prototype to assist the interpretation of machine learning results. Based on the experiments conducted, it can be concluded that student gender shows strong contribution to the dropout prediction than other features. Synthetic Minority Over-Sampling Technique and Edited Nearest Neighbor (SMOTE ENN) data balancing technique provide a good solution for achieving a higher performance. Furthermore, an ensemble classifier which was developed by softly combining the tuned LR2 and MLP2 achieved the best results when SMOTE ENN technique was employed. These results indicate that, the data driven approach and the developed prototype (BakiShule) were capable of accurately predicting and identifying students who are at risk of dropping out of school for early intervention.

5.2 Recommendations

The developed prototype (BakiShule) has correctly identified students and schools with high dropout rates. We therefore encourage schools, teachers, parents, the Ministry of Education, Science and Technology, NGOs and other education stakeholders to consider using BakiShule on identifying students at risk of dropping out and schools with high dropout rate for early interventions and monitoring.

The developed prototype only identifies and predicts students at risk of dropping out of school for early intervention. Other studies can be carried out to use BakiShule as a data collection tool and incorporate the generated data in the system's ability to suggest a strategic intervention for a student or a group of students who are risk of dropping out and point out who among the stakeholders should be responsible to best assist the such students.

During development of this data driven approach, three datasets were used to train the model and the model achieved a test score of 85%. However, to diversify the predictive models' functionalities such as ranking students according to their probability of dropping out of school and increasing the test score for more accurate predictions, other studies are recommended to build upon BakiShule and use more datasets and incorporate cost-benefit analysis and socio-economic factors that were not included in our developed approach due to limited time and cost.

The developed prototype is web-based therefore it requires internet connection to work. However, due to poor and intermittent internet access particularly in remote areas, we recommend other studies to consider developing approaches and/or mobile applications that can work offline and allow users to access and interacts with the system at any time hence reducing the burden of internet cost to users.

The study worked with education stakeholders such as teachers, parents, students, systems development experts and education officers from five districts (Arusha, Kisarawe, Mbeya, Nzega and Rufiji) with high, mid and low dropout rates. Due to the nature of the dropout problem, we recommend the Ministry of Education, Science and Technology, NGOs and other studies to carry out surveys across all districts in the country to come up with more comprehensive information that may help during model development for addressing the student dropout problem in Tanzania.

The results from this study showed that, girls are at higher risk of dropping out of school than boys. We therefore recommend that policies and other approaches which will target this particular group to be urgently formulated.

REFERENCES

Abdi, L., & Hashemi, S. (2014). An Ensemble Pruning Approach Based on Reinforcement Learning in Presence of Multi-class Imbalanced Data. *Proceedings of the Third International Conference on Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing*, 258. https://doi.org/10.1007/978-81-322-1771-8

Afolabi, L. T., Saeed, F., Hashim, H., & Petinrin, O. O. (2018). Ensemble Learning Method for the Prediction of New Bioactive Molecules. *PLoS ONE*, *13*(1), 1–14. https://doi.org/10.1371/journal.pone.0189538

Aguiar, E., Dame, N., Miller, D., Yuhas, B. & Addison, K. L. (2015). Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of not Graduating High School on Time Categories and Subject Descriptors. *ACM*, 93–102.

Allensworth, E., & Easton, J. (2007). What Matters for Staying On-track and Graduating in Chicago Public High Schools. In *Consortium on Chicago School Research* (Issue 1).

Amadi, M. A., Role, E., & Makewa, L. N. (2013). Girl Child Dropout: Experiential Teacher and Student Perceptions Curriculum and Teaching View Project Girl Child Dropout: Experiential Teacher and Student Perceptions. *International Journal of Humanities and Social Science*, *3*(5), 124-131. www.ijhssnet.com

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival Analysis based Framework for Early Prediction of Student Dropouts. *Proceedings of the ACM Conference on Information and Knowledge Management*, 24–28. https://doi.org/10.1145/2983323.2983351

Amin, M. Z., & Ali, A. (2017). *Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions* (Issue 1). https://doi.org/10.13140/RG.2.2.26371.25127

Ananga, E. D. (2011). Typology of School Dropout: The Dimensions and Dynamics of Dropout in Ghana. *International Journal of Educational Development*, *31*(4), 374–381. https://doi.org/10.1016/j.ijedudev.2011.01.006

Archambault, I., Janosz, M., Fallu, J. S. & Pagani, L. S. (2009). Student engagement and its

relationship with early high school dropout. *Journal of Adolescence*, 32, 651–70. https://dx.doi.org/10.1016/j.adolescence.2008.06.007

Arsad, M. P., Buniyamini, N., & Manan, J. A. (2013). A Neural Network Students' Performance Prediction Model (NNSPPM). *Proceeding of the IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, *11*, 1–5. https://doi.org/10.1109/ICSIMA.2013.6717966

Ashimolowo, O. R., Aromolaran, A. K., & Inegbedion, S. O. (2010). Child Street - Trading Activities and Its Effect on the Educational Attainment of Its Victims in Epe Local Government Area of Lagos State. *Journal of Agricultural Science*, *2*(4), 211-220. https://doi.org/10.5539/jas.v2n4p211

Aulck, L., Aras, R., Li, L., Heureux, C. L., Lu, P., & West, J. (2017). STEM-ming the Tide : Predicting STEM Attrition Using Student Transcript Data. *In Proceedings of ACM Knowledge Discovery and Data Mining Conference*, 1–10. https://doi.org/10.475/123_4

Bäckman, O. (2017). High School Dropout, Resource Attainment, and Criminal Convictions. *Journal of Research in Crime and Delinquency*, *54*(5), 715–749. https://doi.org/10.1177/0022427817697441

Baker, W. L. (2011). High School Droput: Perceptions and Voices of African American and Hispanic Students. In *Texas A&M University* (Issue 5).http://ezphost.dur.ac.uk/login?https://search.proquest.com/docview/885654612?accountid=14533%0Ahttp://openurl.ac.uk/ukfed:dur.ac.uk?genre=dissertations+%26+theses&issn=&title=High+school+droput%3A+Perceptions+and+voices+of+African+American+and+Hispanic

Barrat, V. X., Berliner, B. A., & Fong, A. B. (2012). When Dropping Out is Not a Permanent High School Outcome: Student Characteristics, Motivations, and Reenrollment Challenges. *Journal of Education for Students Placed at Risk*, *17*(4), 217–233.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Exploration Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Batuwita, R., & Palade, V. (2012). Adjusted Geometric-mean: A Novel Performance Measure for Imbalanced Bioinformatics Datasets Learning. *Journal of Bioinformatics and*

Computational Biology, *10*(4), 1-23. https://doi.org/10.1142/S0219720012500035

Beck, H. P., & Davidson, W. D. (2001). Establishing an Early Warning System : Predicting Low Grades in College Students from Survey of Academic Orientations Scores. *Research in Higher Education*, *42*(6), 709-721. https://doi.org/10.1023/A

Bell, B., Costa, R., & Machin, S. (2017). Crime-Age Profiles and School Dropout. *Quantitative Social Science*. 1-48. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RESConf2018&paper_id=1062

BEST. (2015). *Pre-Primary, Primary and Secondary Education Statistics in Brief 2016 The United Republic of Tanzania President's Office Regional Administration and Local Government*. https://s-u-west-1.amazonaws.com/s3.sourceafrica.net/documents/118112/Tanzania-Pre-Primary-Primary-and-Secondary.pdf

BEST. (2018). *President's Office Regional Administration and Local Government*. http://www.stat.gov.pl/cps/rde/xchg/gus

Borowska, K., & Topczewska, M. (2016). New Data Level Approach for Imbalanced Data Classification Improvement. *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015, Advances in Intelligent Systems and Computing*, 283–294. https://doi.org/10.1007/978-3-319-26227-7

Bouaguel,W., Mufti, G. B., & Limam, M. (2015). A New Feature Selection Technique Applied to Credit Scoring Data Using a Rank Aggregation Approach Based on: Optimization, Genetic Algorithm and Similarity. *In Knowledge Discovery Process and Methods to Enhance Organizational Performance*. https://doi.org/10.1201/b18231

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do We Know Who Will Drop Out?: A Review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity. *The High School Journal*. https://doi.org/10.1353/hsj.2013.0000

Branson, N., Hofmeyr, C., & Lam, D. (2014). Progress Through School and the Determinants of School Dropout in South Africa. *Development Southern Africa*, *31*(1), 106–126. https://doi.org/10.1080/0376835X.2013.853610

Burez, J., & Van den Poel, D. (2009). Handling Class Imbalance in Customer Churn Prediction.

Expert Systems with Applications, *36*(3), 4626–4636.

Can, E., Oya Aktas, F., & Tuzun Arpacioglu, I. (2017). The Reasons of School Dropouts in Higher Education: Babaeski Vocational College Case. *Universal Journal of Educational Research*. 5(12), 84–88. http://www.hrpub.org/journals/article info.php?aid=6624

Center for Digital Technology and Management. (2015). *The Future of Education Trend*.1-123. https://issuu.com/cdtm/docs/the_future_of_education_-_cdtm_tren

Chen, R. (2012). Institutional Characteristics and College Student Dropout Risks: A Multilevel Event History Analysis. *Research in Higher Education.* 53, 487–505. https://doi.org/10.1007/s11162-011-9241-4

Chung, J. Y., Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*. 96, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Ciolacu, M., Svasta, P. M., Berg, W., & Popp, H. (2017). Education for Tall Thin Engineer in a Data Driven Society. *International Symposium for Design and Technology in Electronic Packaging*, *1*, 432–437. https://doi.org/10.1109/SIITME.2017.8259942

Development Education Research Centre. (2018). *Global Education Digest Report*. www.ucl.ac.uk/ioe-derc

Dockery, D. J. (2012). *School Dropout Indicators, Trends, and Interventions for School Counselors Donna J. Dockery Virginia Commonwealth University*.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting Student Performance Using Personalized Analytics. *Computer*, *49*(4), 61–69. https://doi.org/10.1109/MC.2016.119

Elhassan, T., Aljurf, M., & Shoukri, M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Journal of Informatics and Data Mining*, *1*(2), 1–12.

Estêvão, P., & Álvares, M. (2014). What do we Mean by School Dropout? Early School Leaving and the Shifting of Paradigms in School Dropout Measurement. *Portuguese Journal of Social Science*, *13*(1), 21–32. https://doi.org/10.1386/pjss.13.1.21_1

Fall, A. M., & Roberts, G. (2012). High School Dropouts: Interactions between Social Context, Self-perceptions, School Engagement, and Student Dropout. *Journal of Adolescence*, 35(4), 787-798. http://doi:10.1016/j.adolescence.2011.11.004

Farah, N., & Upadhyay, M. P. (2017). How are School Dropouts Related to Household Characteristics? Analysis of Survey Data from Bangladesh. *Cogent Economics and Finance*. 5(1),1–18. http://dx.doi.org/10.1080/23322039.2016.1268746

Farquad, M. A. H., & Bose, I. (2012). Preprocessing Unbalanced Data Using Support Vector Machine. *Decision Support Systems*, *53*(1), 226–233.

Fei, M., & Yeung, D. Y. (2015). Temporal Models for Predicting Student Dropout in Massive Open Online Courses. *IEEE International Conference on Data Mining Workshop (ICDMW)*, 256–263. https://doi.org/10.1109/ICDMW.2015.174

Funja, R. (2018). *Fighting School Dropouts Among Girls*. https://dcli.co/impact-story/fighting-school-dropouts-among-girls/

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2016). New Ordering-Based Pruning Metrics for Ensembles of Classifiers in Imbalanced Datasets. *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015, Advances in Intelligent Systems and Computing*, 1–13.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). *How do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm* (Issue 834). https://www.bis.org/publ/work834.pdf

Gao, T. (2015). *Hybrid Classification Approach of SMOTE and Instance Selection for Imbalanced Datasets*. Iowa State University. https://pdfs.semanticscholar.org/7d1f/f09cd0b23305b33e0f1173ed0c9300c9aba0.pdf

Gray, G., McGuinness, C., & Owende, P. (2014). An Application of Classification Models to Predict Learner Progression in Tertiary Education. *IEEE International Advance Computing Conference, IACC*, 549–554. https://doi.org/10.1109/IAdCC.2014.6779384

Habibipour, A., Georges, A., Ståhlbröst, A., Schuurman, D., & Bergvall-Kåreborn, B. (2018). A Taxonomy of Factors Influencing Drop-Out Behaviour in Living Lab Field Tests.

Technology Innovation Management Review, *8*(5), 5–21. https://doi.org/10.22215/timreview/1155

Hailikari, T., Nevgi, A., & Lindblom-Ylänne, S. (2007). Exploring Alternative Ways of Assessing Prior Knowledge, its Components and their Relation to Student Achievement: A Mathematics Based Case Study. *Studies in Educational Evaluation*, *33*(3–4), 320–337. https://doi.org/10.1016/j.stueduc.2007.07.007

Hakim, A. (2019). *Performance Evaluation of Machine Learning Techniques for Early Prediction of Brain Strokes*. http://dspace.uiu.ac.bd/bitstream/handle/52243/1508/Md%20Azizul%20Hakim%20MSCSE%20ID%20012182003.pdf?sequence=1&isAllowed=y

Halland, R., Igel, C. & Alstrup, S. (2015). High-School Dropout Prediction Using Machine Learning : A Danish Large-scale Study. *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 22–24.

Herzog, S. (2006). Estimating student retention and degree-completion time: decision trees and neural networks vs regression. *New Directions for Institutional Research*. 131, 17–33

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced Datasets: From Sampling to Classifiers. In *Imbalanced Learning: Foundations, Algorithms, and Applications* (pp. 43–59). https://doi.org/10.1002/9781118646106.ch3

Human Right Watch. (2017). *"I Had a Dream to Finish School":Barriers to Secondary Education in Tanzania*.https://www.hrw.org/sites/default/files/report_pdf/tanzania0217_insert_lowres_spreads.pdf

Hunt, X. J., Carolina, N., Carolina, N., Silva, J., & Carolina, N. (2017). Transfer Learning for Education Data. *Proceedings of ACM SIGKDD Conference, El Halifax, Nova Scotia Canada*, 1–6.

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using Machine Learning to Predict Student Difficulties from Learning Session Data. *Artificial Intelligence Review*, *52*(1), 381–407. https://doi.org/10.1007/s10462-018-9620-8

Iam-On, N., & Boongoen, T. (2017). Generating Descriptive Model for Student Dropout: A

Review of Clustering Approach. *Human-centric Computing and Information Sciences*. 7(1). http://hcis-journal.springeropen.com/articles/10.1186/s13673-016-0083-0

Injendi, J. (2014). *Secondary School Students' Dropouts And Crime Escalation In Vihiga County, Kenya*.

Isphording, I., & Qendrai, P. (2019). Gender Differences in Student Dropout in STEM. *Journal of Family Research*, *3*(82), 126–134. https://doi.org/10.5157/NEPS

Johnson, R. A., Dame, N., Greatorex-voith, S., & Fritzler, A. (2015). A Data-Driven Framework for Identifying High School Students at Risk of Not Graduating on Time. *Bloomberg Data for Good Exchange Conference*. https://www3.nd.edu/~dial/publications/johnson2015data.pdf

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives and Prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joseph, H. R. (2014). Promoting Education: A state of the Art Machine Learning Framework for Feedback and Monitoring e-Learning Impact. *IEEE Global Humanitarian Technology Conference - South Asia Satellite, GHTC-SAS 2014*, 251–254.

Kassam, Y. (2000). Julius Kambarage Nyerere. *Prospects*, *24*(1–2), 247–259. https://doi.org/10.1007/bf02199019

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer Credit Risk Models via Machine Learning Algorithms. *Social Science Research Network*, 1–49. https://doi.org/10.2139/ssrn.1568864

Kim, A., Song, Y., Kim, M., Lee, K., & Cheon, J. H. (2018). Logistic Regression Model Training based on the Approximate Homomorphic Encryption. *International Association for Cryptologic Research Archive*, 1–13. https://eprint.iacr.org/2018/254

Kim, M. J., Kang, D. K., & Kim, H. B. (2015). Geometric Mean-based Boosting Algorithm with Over-sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction. *Expert Systems with Applications*, *42*(3), 1074–1082. https://doi.org/10.1016/j.eswa.2014.08.025

Kim S. (2019). Data-based Decision-making for School Improvement: Research Insights and

Gaps. *Educational Research*, 61(3), 257-273. https://10.1080/00131881.2019.1625716

Kotsiantis, S. B. (2012). Use of Machine Learning Techniques for Educational Proposes: A Decision Support System for Forecasting Students' Grades. *Artificial Intelligence Review*, *37*(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. *Springer*, 267–274.

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krstic, K., Stepanovic-Ilic, I., & Videnovic, M. (2017). Student Dropout in Primary and Secondary Education in the Republic of Serbia. *Psiholoska Istrazivanja*, *20*(1), 27–50. https://doi.org/10.5937/psistra1701027k

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine Learning in Education - A Survey of Current Research Trends. *Annals of DAAAM and Proceedings of the International DAAAM Symposium*, *29*(1), 406–410. https://doi.org/10.2507/29th.daaam.proceedings.

Kumar, M., Singh, A., & Handa, D. (2017). Literature Survey on Educational Dropout Prediction. *International Journal of Education and Management Engineering*. 7(2), 8–19. http://www.mecs-press.org/ijeme/ijeme-v7-n2/v7n2-2.html

Kuncheva, L. I., Arnaiz-González, Á., Díez-Pastor, J. F., & Gunn, I. A. D. (2019). Instance Selection Improves Geometric Mean Accuracy: A Study on Imbalanced Data Classification. *Progress in Artificial Intelligence*, *8*(2), 215–228. https://doi.org/10.1007/s13748-019-00172-4

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909–1918. https://doi.org/10.1145/2783258.2788620

Latif, A., Choudhary, A., & Hammayun, A. (2015). Economic Effects of Student Dropouts: A

Comparative Study. *Journal of Global Economics*, *03*(02), 2–5. https://doi.org/10.4172/2375-4389.1000137

Lekwa, E. A., & Anyaogu, B. E. (2016). Economic Recession, Hawking and Students Dropout of School in the Five Eastern States of Nigeria. *International Journal of Social Sciences and Management Research*, *2*(2), 14–21. www.iiardpub.org

Lewin, K. M. (2009). Access to Education in Sub-Saharan Africa: Patterns, Problems and Possibilities. *Comparative Education*, *45*(2), 151–174.

Liang, J., Li, C., & Zheng, L. (2016). Machine Learning Application in MOOCs: Dropout Prediction. *International Conference on Computer Science and Education*. 52–57.

Lin, W. J., & Chen, J. J. (2013). Class-imbalanced Classifiers for High-dimensional Data. *Briefings in Bioinformatics*, *14*(1), 13–26. https://doi.org/10.1093/bib/bbs006

Lockett, C., & Cornelious, L. (2015). Factors Contributing to Secondary School Dropouts in an Urban School District. *Research in Higher Education Journal*, *29*, 1–15. http://www.aabri.com/manuscripts/152331.pdf

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network*, *2*(1), 83–87. https://doi.org/10.1109/SIU.2013.6531574

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An Insight into Classification with Imbalanced Data : Empirical Results and Current Trends on using Data Intrinsic Characteristics. *Information Sciences*, *250*, 113–141.

Ma, C., Yao, B., Ge, F., Pan, Y., & Guo, Y. (2017). Improving Prediction of Student Performance Based on Multiple Feature Selection Approaches. *Proceedings of the International Conference on E-Education, E-Business and E-Technology*, 36–41. http://doi.acm.org/10.1145/3141151.3141160

Maginga, T. J., Nordey, T., & Ally, M. (2018). Extension System for Improving the Management of Vegetable Cropping Systems. *Journal of Information Systems Engineering & Management, 3*(4). https://doi.org/10.20897/jisem/3940

Matto, G. J., & Mwangoka, J. (2018). Mining Frequent Patterns of Crime using FP-Growth

with Multiple Minimum Supports based on Shannon Entropy. *International Journal of Computer Applications,* 180(24).

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa, F. H., & Ventura, S. (2016). Early Dropout Prediction using Data Mining: A Case Study with High School Students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Mazumder, R. U., Begum, S. A., & Biswas, D. (2015). Rough Fuzzy Classification for Class Imbalanced Data. *Proceedings of Fourth International Conference on Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing*, 159–171. https://doi.org/10.1007/978-81-322-2217-0

Mduma, N., Kalegele, K., & Machuve, D. (2019). Machine learning approach for reducing students dropout rates. *International Journal of Advanced Computer Research, 9*(42), 156-169. https://doi.org/10.19101/IJACR.2018.839045

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal*, *18*, 1–10. https://doi.org/10.5334/dsj-2019-014

Mgala, M. (2016). *Investigating Prediction Modelling of Academic Performance for Students in Rural Schools in Kenya*. University of Cape Town.

Mgala, M., & Mbogho, A. (2015). Data-driven Intervention-level Prediction Modeling for Academic Performance. *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, 1–8. https://doi.org/10.1145/2737856.2738012

Ministry of Education. (2015). *The United Republic of Tanzania Education Sector Development* (Vol. 2). https://www.globalpartnership.org/sites/default/files/2019-04-gpe-tanzania-esp.pdf

Moore, A. (2017). Factors That Cause Students to Leave Before Graduation. In *Carson-Newman University*. https://doi.org/10.1016/j.sbspro.2015.04.758

Morara, A. N., & Chemwei, B. (2013). Dropout among Pupils in Rural Primary Schools in Kenya : The Case of Nandi North District, Kenya. *Journal of Education and Practice*,

4(19), 1–13.

Morris, L. V., Wu, S. S., & Finnegan, C. L. (2005). Predicting retention in online general education courses. *American Journal of Distance Education*. 19(1), 23–36.

Mosha, D. (2014). *Assessment of Factors behind Dropout in Secondary Schools in Tanzania. A Case of Meru District in Tanzania*. Open University of Tanzania.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Murray, M. (2014). Factors Affecting Graduation and Student Dropout Rates at the University of KwaZulu-Natal. *South African Journal of Science*, *110*(11–12), 1–6. https://doi.org/10.1590/sajs.2014/20140008

Nacheva, R. (2017). Prototyping Approach in User Interface. *2nd Conference on Innovative Teaching Methods*, *June*, 80–87. https://www.researchgate.net/publication/317414969

Nakpodia, E. D. (2010). An Analysis of Dropout Rate among Secondary School Student in Delta State, Nigeria (1999-2005). *Journal of Social Sciences*, *23*(2), 99–103. https://doi.org/10.1080/09718923.2010.11892817

Natek, S., & Zwilling, M. (2014). Student Data Mining Solution – Knowledge Management System Related to Higher Education Institutions. *Expert Systems with Applications*, *41*, 6400–6407. https://doi.org/10.1016/j.eswa.2014.04.024

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An Early Warning System. *Educational Leadership*, *October*, 28–33.

Neill, A., Yang, D., Roy, M., Sebastiamphillai, S., Hofer, S., & Xu, W. (2020). Development and Evaluation of a Machine Learning Prediction Model for Flap Failure in Microvascular Breast Reconstruction. *Annals of Surgical Oncology*. https://doi.org/10.1245/s10434-020-08307-x

Neumann, P. (2004). Prototyping. In *Topic Report* (pp. 1–13). http://pages.cpsc.ucalgary.ca/~saul/pmwiki/uploads/Main/topic-neumann.pdf

Nielsen, K. (2016). Engagement, Conduct of Life and Dropouts in the Danish Vocational

Education and Training (VET) System. *Journal of Vocational Education & Training*, *68*(2), 198–213.

OECD. (2015). *Education Policy Outlook 2015: Making Reforms Happen* (Issue 11). https://www.oecd.org/edu/Japan-country-profile.pdf

Oruko, K., Nyothach, E., Zielinski-Gutierrez, E., Mason, L., Alexander, K., Vulule, J., Laserson, K. F., & Phillips-Howard, P. A. (2015). He is the One Who is Providing you with Everything so Whatever he Says is What you Do: A Qualitative Study on Factors Affecting Secondary Schoolgirls' Dropout in Rural Western Kenya. *PLoS ONE*, *10*(12), 1–14. https://doi.org/10.1371/journal.pone.0144321

Otieno, G. C. (2016). *Socio-Economic Factors Influencing Students' Dropout Rates In Public Secondary Schools In Msambweni Sub-County, Kwale County, Kenya*. University of Nairobi.

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial Intelligence, Machine Learning and Health Systems. *Journal of Global Health*, *8*(2), 1-8. https://doi.org/10.7189/jogh.08.020303

Prieto, L. P., Rodr´ıguez-Triana, M. J., Kusmin, M. & Laanpere, M. (2017). Smart School Multimodal Dataset and Challenges. *CEUR Workshop Proceedings*. 1828, 53–59.

Prusa, J., Khoshgoftaar, T. M., DIttman, D. J., & Napolitano, A. (2015). Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. *IEEE 16th International Conference on Information Reuse and Integration, IRI 2015*, 197–202. https://doi.org/10.1109/IRI.2015.39

Rakesh, A., Christoforaki, M., Gollapudi, S., Kannan, A., Kenthapad, K., & Swaminathan, A. (2014). Mining Videos from Web for Electronic Textbooks. *Microsoft Research*. 8478, 219-231.

Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). SMOTE-RSB *: A Hybrid Preprocessing Approach Based on Oversampling and Undersampling for High Imbalanced Data-sets Using SMOTE and Rough Sets Theory. *Knowledge and Information Systems*, *33*(2), 245–265. https://doi.org/10.1007/s10115-011-0465-6

Rannveig, S. (2016). *Academic Performance and Student Dropout*. https://doi.org/10.1016/0168-9525(88)90171-0

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The Importance of Prior Knowledge When Comparing Examples: Influences on Conceptual and Procedural Knowledge of Equation Solving. *Journal of Educational Psychology*, *101*(4), 836 -852.

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven System to Predict Academic Grades and Dropout. *PLoS ONE*, *12*(2), 1–21. https://doi.org/10.1371/journal.pone.0171207

Rumberger, R. W., Addis, H., Allensworth, E., Balfanz, R., Duardo, D., & Dynarski, M. (2017). Preventing Dropout in Secondary Schools. In *National Center for Educational Evaluation and Regional Assistance*. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/wwc_dropout_092617.pdf%0AAll Papers/R/Rumberger *et al.* 2017 - Preventing Dropout in Secondary Schools.pdf

Rumberger, R. W., & Lim, S. A. (2008). Why Students Drop Out of School: A Review of 25 Years of Research. In *California Dropout Research Project Report*. https://www.issuelab.org/resources/11658/11658.pdf

Rutakinikwa, L. N. (2016). *Factors Influencing Secondary School Girls Dropout in Bagamoyo*. https://www.google.com/search?bm&source=hp&ei=p7HgXterHIbVgQbEq6mYBw&q

Santana, M. A., Costa, E. B., Neto, B. F. S., Silva, I. C. L., & Rego, J. B. A. (2015). A Predictive Model for Identifying Students with Dropout Profiles in Online Courses. *CEUR Workshop Proceedings*. https://pdfs.semanticscholar.org/.pdf

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-School Dropout Prediction Using Machine Learning : A Danish Large-scale Study. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, *4*, 22–24.

Sathya, R., & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38. https://doi.org/10.14569/IJARAI.2013.020206

Shahidul, S. M., & Karim, A. H. (2015). Factors Contributing to School Dropout among the Girls: A review of literature. *European Journal of Research and Reflection in Educational*

Sciences, *3*(2), 25–36. https://www.idpublications.org/wp-content/uploads/2015/02

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shilbayeh, S. A. (2015). *Cost Sensitive Meta Learning*. http://usir.salford.ac.uk/id/eprint/36278/

Simic, N., & Krstic, K. (2017). School Factors Related to Dropout from Primary and Secondary Education in Serbia: A Qualitative Research. *Psiholoska istrazivanja*. 20(1), 51–70. http://scindeks.ceon.rs/Article.aspx?artid=0352-73791701051S

Stempel, H., Cox-Martin, M., Bronsert, M., Dickinson, L. M., & Allison, M. A. (2017). Chronic School Absenteeism and the Role of Adverse Childhood Experiences. *Academic Pediatrics*, *17*(8), 837–843. https://doi.org/10.1016/j.acap.2017.09.013

Sun, S., Zhang, C., & Zhang, Y. (2017). Traffic Flow Forecasting Using a Spatio - Temporal Bayesian Network Predictor. *Proceedings of International Conference of Artificial Neural Networks: Formal Models and Their Applications*. https://link.springer.com/chapter/10.1007/11550907_43

Tabuchi, T., Fujihara, S., Shinozaki, T., & Fukuhara, H. (2018). Determinants of High-School Dropout: A Longitudinal Study in a Deprived Area of Japan. *Journal of Epidemiology*, 14–16. https://doi.org/10.2188/jea.JE20170163

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition. *Expert Systems with Applications*, *41*(2), 321–330. https://doi.org/10.1016/j.eswa.2013.07.046

Trevor, C., Marshall, A., & Allen, C. (2018). *Guide to DHS Statistics* (Issue 9). http://www.measuredhs.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_29Oct2012_DHSG1.pdf%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?

Tufi, S. M., Neimar, D. S. F., N´obrega, M. C., & Nicolella, A. C. (2015). Factors Associated with Dropout Rates in Public Secondary Education in Minas Gerais. *Educação e*

Pesquisa. 41(3), 757–772. https://doi.org/10.1590/S1517-9702201507138589

UNESCO. (2011). *UNESCO Global Partnership for Girls' and Women's Education- One Year On*. http://www.unesco.org/eri/cp/factsheets_ed/TZ_EDFactSheet.pdf

UNESCO. (2017). *Estimation of the Numbers and Rates of Out-of-school Children and Adolescents Using Administrative and Household Survey Data* (Issue 4).

United Republic of Tanzania. (2015). *The National Education Act, 2015*. http://www.ilo.org/dyn/natlex/natlex4.detail?p_lang=en&p_isn=94089&p_country=T&p _count=270

Vaishnavi, V., & Kuechler, W. (2015). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology* (1st ed.). Auerbach Publications.

Virtanen, T., & Tuomo. (2016). Student Engagement in Finish Lower Secondary School [University of Jyväskylä]. In *Jyväskylä Studies in Education, Psychology and Social Research*. https://jyx.jyu.fi/dspace/handle/123456789/51563

Vital Wave Consulting. (2009). mHealth for Development: The Opportunity of Mobile Technology for Healthcare in the Developing World. *Technology*, *46*(1), 1–70. https://doi.org/10.1145/602421.602423

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., & Wollscheid, S. (2015). Drop-out and Completion in Higher Education in Europe. In *European Commission*. https://doi.org/10.2766/826962

Wang, S., & Yao, X. (2013). Using Class Imbalance Learning for Software Defect Prediction. *IEEE Transactions on Reliability*, *62*(2), 434–443. https://doi.org/10.1109/TR.2013.2259203

Warfel, T. (2009). *Prototyping: A Practitioner's Guide* (M. Justak (ed.); 1st ed.). Rosenfeld Media. https://www.worldcat.org/title/prototyping/oclc/775301615?referer=&ht=edition

Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse Factor Analysis for Learning and Content Analytics. *Journal of Machine Learning Research*, *15*, 1959–2008.

Weiss, G., McCarthy, K., & Zabar, B. (2007). Cost-Sensitive Learning vs. Sampling: which is Best for Handling Unbalanced Classes with Unequal Error Costs?. *Proceedings of*

International Conference in Data Mining, pp. 35-41. https://www.semanticscholar.org/paper/Cost-Sensitive-Learning-vs.-Sampling%3A-Which-is-Best-Weiss McCarthy/9ebd

Willging, P. A., & Johnson, S. D. (2009). Factors that Influence Students' Decision to Dropout of Online Courses. *Journal of Asynchronous Learning Network*, *13*(3), 115–127. https://doi.org/10.24059/olj.v8i4.1814

Wu, W., May, R., Maier, H., & Dandy, G. C. (2013). A Benchmarking Approach for Comparing Data Splitting Methods for Modeling Water Resources Parameters Using Artificial Neural Networks. *Water Resources Research*, *4*9, 7598–7614. https://doi.org/10.1002/2012WR012713

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine Learning in Manufacturing: Advantages, Challenges and Applications. *Production and Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Xu, J., Moon, K. H., & van der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742–753. https://doi.org/10.1109/ JSTSP.2017.2692560

Yen, S. J., & Lee, Y. S. (2009). Cluster-based Under-sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*, *36*(3), 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (1st ed., pp. 1–10). Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-39112-5_18

Zheng, S., Rosson, M., Shih, P., & Carroll, J. (2015). Understanding Student Motivation, Behaviors and Perceptions in MOOCs. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1882–1895.

Zubair, H. K. M., & Zahid, H. M. (2019). Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease. *Advances in Intelligent Systems and Computing*, *882*, 415–426. https://doi.org/10.1007/978-981-13-5953-8_34

APPENDICES

Appendix 1: Round Table and Focus Group Discussion Guiding Questions

1. How do you know the status of student/school dropout?

2. How do you anticipate possible occurrence of dropout problem?

3. Do you have technological tool/system to predict dropout and prevent them?

4. Do you make use of available data to anticipate dropout occurrence?

5. What are the factors contributed to student dropout problem?

6. Discuss the process flow of planning and budgeting in Tanzania?

Appendix 2: Evaluation Questionnaires

(i) Prototype Technical Evaluation Questionnaire

Evaluation Criteria	Very High	High	Average	Low	Very Low
Ability of the system to be accessed					
Scalability of the system					
User friendliness of the system					
Consistency of the system					
Ability of the system to be navigated					
Ability of the system to give feedback					

(ii) Prototype End-user Evaluation Questionnaire

Evaluation Criteria	Very High	High	Average	Low	Very Low
Ability of the system to predict student dropout					
Ability of the system to visualize school dropout					
Clarify of the predicted results					
Usefulness of the system					

(iii) Fomu ya kutathmini mfumo wa BakiShule kwa wataalamu

Kigezo cha kutathmini	**Vizuri sana**	**Vizuri**	**Wastani**	**Chini (Sababu)**	**Chini sana (Sababu)**
Urahisi wa kupatikana					
Uwezo wa kutanuka					
Urahisi wa kutumia					
Consistency of the system					
Urahisi wa kuhama kurasa					
Uwezo wa kurudisha majibu					

(iv) Fomu ya kutathmini mfumo wa BakiShule kwa watumiaji

Kigezo cha kutathmini	**Vizuri sana**	**Vizuri**	**Wastani**	**Chini (Sababu)**	**Chini sana (Sababu)**
Uwezo wa mfumo kutabiri					
Uwezo wa mfumo kuonyesha anguko la mwanafunzi					
Usawa wa majibu ya utabiri					
Umuhimu wa BakiShule					

Appendix 3: Codes for Feature Engineering

```
from numpy.random import seed
seed(7)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn-white')
from pylab import rcParams
%matplotlib inline
np.set_printoptions(precision=3)
dark_colors = ["#A51C30", "#808080",
(0.8509803921568627, 0.37254901960784315, 0.00784313725490196),
(0.4588235294117647, 0.4392156862745098, 0.7019607843137254),
(0.9058823529411765, 0.1607843137254902, 0.5411764705882353),
(0.4, 0.6509803921568628, 0.11764705882352941),
(0.9019607843137255, 0.6705882352941176, 0.00784313725490196),
(0.6509803921568628, 0.4627450980392157, 0.11372549019607843),
(0.4, 0.4, 0.4)]
rcParams['figure.dpi'] = 600
rcParams['axes.color_cycle'] = dark_colors
rcParams['axes.facecolor'] = "white"
rcParams['patch.edgecolor'] = 'none'
rcParams['grid.color']="gray"
rcParams['grid.linestyle']="-"
rcParams['grid.linewidth'] = 0.3
rcParams['grid.alpha']=1
rcParams['text.color'] = "444444"
rcParams['axes.labelcolor'] = "444444"
rcParams['ytick.color'] = "444444"
rcParams['xtick.color'] = "444444"
from sklearn import preprocessing
```

```
from imblearn.over_sampling import SMOTE
from imblearn.combine import SMOTEENN
from imblearn.under_sampling import RandomUnderSampler,
RepeatedEditedNearestNeighbours
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier, from
sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB, BernoulliNB, MultinomialNB
from sklearn.linear_model import SGDClassifier
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis as from
sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from sklearn import svm
from sklearn import model_selection, cross_validation
from sklearn.metrics import classification_report, confusion_matrix, roc_curve, from
sklearn.cross_validation import StratifiedShuffleSplit
from sklearn.utils import shuffle
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import VotingClassifier
from math import sqrt
import matplotlib as mpl
import matplotlib
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from pylab import rcParams
SPINE_COLOR = 'gray'
def latexify(fig_width=None, fig_height=None, columns=1):
"""Set up matplotlib's RC params for LaTeX plotting.
Call this before plotting a figure.
Parameters
----------
```

```
fig_width : float, optional, inches
fig_height : float, optional, inches
columns : {1, 2}
"""
# code adapted from http://www.scipy.org/Cookbook/Matplotlib/LaTeX_Examples
# Width and max height in inches for IEEE journals taken from
#
computer.org/cms/Computer.org/Journal%20templates/transactions_art_guide.assert(columns
in [1,2])
if fig_width is None:
fig_width = 3.39 if columns==1 else 6.9 # width in inches
if fig_height is None:
golden_mean = (sqrt(5)-1.0)/2.0 # Aesthetic ratio
fig_height = fig_width*golden_mean # height in inches
MAX_HEIGHT_INCHES = 8.0
if fig_height > MAX_HEIGHT_INCHES:
print("WARNING: fig_height too large:" + fig_height +
"so will reduce to" + MAX_HEIGHT_INCHES + "inches.")
fig_height = MAX_HEIGHT_INCHES
params = {'backend': 'ps',
'text.latex.preamble': ['\\usepackage{gensymb}'],
'axes.labelsize': 10, # fontsize for x and y labels (was 10)
'axes.titlesize': 10,
'font.size': 10, # was 10
'legend.fontsize': 10, # was 10
'xtick.labelsize': 10,
'ytick.labelsize': 10,
'text.usetex': True,
'axes.titlepad': 20,
'figure.figsize': [fig_width,fig_height],
'font.family': 'serif'
}
matplotlib.rcParams.update(params)
def format_axes(ax):
```

```
for spine in ['top', 'right']:
ax.spines[spine].set_visible(False)
for spine in ['left', 'bottom']:
ax.spines[spine].set_color(SPINE_COLOR)
ax.spines[spine].set_linewidth(0.5)
ax.xaxis.set_ticks_position('bottom')
ax.yaxis.set_ticks_position('left')
for axis in [ax.xaxis, ax.yaxis]:
axis.set_tick_params(direction='out', color=SPINE_COLOR)
# matplotlib.pyplot.tight_layout()
return ax
def handle_imbalance(data, type="SMOTE"):
y = data.loc[:,'Dropout']
X = data.drop('Dropout',1)
if type == "SMOTE":
smote = SMOTE(kind = "regular")
X, y = smote.fit_sample(X, y)
elif type =="SMOTEENN":
smote_enn = SMOTEENN(random_state=0)
X, y = smote_enn.fit_sample(X, y)
elif type=="undersample":
rus = RandomUnderSampler(random_state=0)
X, y = rus.fit_sample(X, y)
elif type == "rpt" :
rpt=RepeatedEditedNearestNeighbours(random_state=0)
X, y = rpt.fit_sample(X, y)
else:
print("No such sampling techniques")
return X, y
def load_cleandata(file_name):
'''
Function to load saved clean data (train, test and valiadtion set)
file_name = path of the data
'''
```

```
df = pd.read_csv(file_name)
df.drop('Unnamed: 0', axis=1, inplace=True)
return df
# load saved training set
train_data = load_cleandata("../data/train.csv")
train_data.columns
old_names = ['$a', '$b', '$c', '$d', '$e']
new_names = ['a', 'b', 'c', 'd', 'e']
df.rename(columns=dict(zip(old_names, new_names)), inplace=True)
X = train_data.drop('Dropout',axis=1)
def get_data(data):
features = ['Studentsex','villmtaaname','Student_age','PTR','PCR',
'PTR','BPLR','PTMR','ParentCheckChildbook','mealsperday','hh_size']
X = data[features]
y = data.Dropout
return X, y
def imbalance_metrics(ypred, ytrue):
confusion= confusion_matrix(ytrue, ypred)
TP = confusion[1, 1]
TN = confusion[0, 0]
FP = confusion[0, 1]
FN = confusion[1, 0]
tpr = TP / float(FN + TP)
tnr = TN / float(TN + FP)
fpr = FP / float(TN + FP)
fnr = FN / float(TP + FN)
ppv = TP / float(TP + FP)
gm = np.sqrt(tpr*tnr)
if tpr > 0:
agm = (gm + tnr)*(FP + TN)/(1 + FP + TN)
else:
agm = 0
auc = (1 + tpr - fpr)/2
f_m = 2*ppv*tnr/(ppv + tnr)
```

```
return gm
#return [gm, agm, auc, f_m]
def test_result(X_val,y_val, model):
ypred_val = model.predict(X_val)
gm, agm, auc, fscore = imbalance_metrics(ypred_val, y_val)
#print('Val-sensitivity:',str(sens))
#print('Val-specifity:',str(spec))
#print('Val-precision:',str(prec))
return gm, agm, auc, fscore
from sklearn.cross_validation import ShuffleSplit
from sklearn.metrics import r2_score
from collections import defaultdict
train_data = load_cleandata("../data/train.csv")
val_data = load_cleandata("../data/validate.csv")
feature = ['id_districtname', 'id_regionname', 'villmtaaname', 'hh_size',
'mealsperday', 'Student_age', 'SchoolhasGirlsroom', 'PTR',
'PCR', 'GPLR', 'BPLR', 'PTMR', 'Student_readbookwithparent',
'Parentteacherdiscussion', 'ParentCheckChildbook', 'EA_area',
'Studentsex', 'Income__source']
names = feature
feature.append('Dropout')
X, Y = handle_imbalance(train_data[feature], type="SMOTEENN")
log = LogisticRegression(class_weight=None, penalty='l1',fit_intercept=C=0.001, tol=1,
random_state=1)
scores = defaultdict(list)
for train_idx, test_idx in ShuffleSplit(len(X), 100, .3):
X_train, X_test = X[train_idx], X[test_idx]
Y_train, Y_test = Y[train_idx], Y[test_idx]
lr = log.fit(X_train, Y_train)
acc= imbalance_metrics(lr.predict(X_test), Y_test)
for i in range(X.shape[1]):
X_t = X_test.copy()
np.random.shuffle(X_t[:, i])
shuff_acc = imbalance_metrics(lr.predict(X_t), Y_test)
```

```
scores[names[i]].append((acc-shuff_acc)*100/acc)
print ("Features sorted by their score:")
print (sorted([(round(np.mean(score), 3), feat) for
feat, score in scores.items()], reverse=True))
results = sorted([(np.around(np.mean(score), decimals=3), feat) for
feat, score in scores.items()], reverse=True)
feature = ['Studentsex','villmtaaname','Student_age','PTR','PCR',
'GPLR','BPLR','PTMR','ParentCheckChildbook','mealsperday','hh_size']
features = [val[1] for val in results]
features label=["Sex", "PCCB", "MLPD", "SPB","PTD","Age", "Village",
"EAarea","HHsize","plot_name="Permutation_features"
latexify(5)
fig = plt.figure()
ax = fig.add_subplot(111)
plt.bar(range(len(results)), [val[0] for val in results], align='center')
#plt.xticks(range(len(results)), [val[1] for val in results])
plt.xticks(range(len(results)), [val for val in label])
plt.xticks(rotation=90)
plt.title("Feature Importance")
plt.ylabel("Score ($\%$)")
format_axes(ax);
plt.savefig('../image/%s.pdf' %(plot_name), format='pdf', bbox_inches='tight', from
sklearn.preprocessing import StandardScaler, Normalizer, RobustScaler
train_data = load_cleandata("../data/train.csv")
val_data = load_cleandata("../data/validate.csv")
feature = ['id_districtname', 'id_regionname', 'villmtaaname', 'hh_size',
'mealsperday', 'Student_age', 'SchoolhasGirlsroom', 'PTR',
'PCR', 'GPLR', 'BPLR', 'PTMR', 'Student_readbookwithparent',
'Parentteacherdiscussion', 'ParentCheckChildbook', 'EA_area',
'Studentsex', 'Income__source']
names = feature
feature.append('Dropout')
X, Y = handle_imbalance(train_data[feature], type="SMOTEENN")
N = StandardScaler()
```

```
N.fit(X)
X = N.transform(X)
log = RandomForestClassifier()
scores = defaultdict(list)
for train_idx, test_idx in ShuffleSplit(len(X), 100, .3):
X_train, X_test = X[train_idx], X[test_idx]
Y_train, Y_test = Y[train_idx], Y[test_idx]
lr = log.fit(X_train, Y_train)
acc= imbalance_metrics(lr.predict(X_test), Y_test)
for i in range(X.shape[1]):
X_t = X_test.copy()
np.random.shuffle(X_t[:, i])
shuff_acc = imbalance_metrics(lr.predict(X_t), Y_test)
scores[names[i]].append((acc-shuff_acc)/acc)
print ("Features sorted by their score:")
print (sorted([(round(np.mean(score), 3), feat) for
feat, score in scores.items()], reverse=True))
results = sorted([(np.around(np.mean(score)*100, decimals=3), feat) for
feat, score in scores.items()], reverse=True)
plt.bar(range(len(results)), [val[0] for val in results], align='center')
plt.xticks(range(len(results)), [val[1] for val in results])
plt.xticks(rotation=90)
```

Appendix 4: Codes for Model Deployment

```
from sklearn.linear_model import LogisticReg= ression
from sklearn.metrics import accuracy_score, classification_report, confusion_import pandas
as pd
import numpy as np
from matplotlib import pyplot as plt
from google.colab import files
dataset = files.upload()
dataset_test = files.upload()
import io
data_train = pd.read_csv(io.BytesIO(dataset['train.csv']))
data_test = pd.read_csv(io.BytesIO(dataset_test['test.csv']))
data_train.head()
data_test.head()
df_meals_train = data_train['mealsperday']
df_meals_test = data_test['mealsperday']
df_age_train = data_train['Student_age']
df_age_test = data_test['Student_age']
df_readbook_train = data_train['Student_readbookwithparent']
df_readbook_test =data_test['Student_readbookwithparent']
df_parent_discussion_train=data_train['Parentteacherdiscussion']
df_parent_discussion_test = data_test['Parentteacherdiscussion']
df_parent_checkbook_train = data_train['ParentCheckChildbook']
df_parent_checkbook_test = data_test['ParentCheckChildbook']
df_gender_train = data_train['Studentsex']
df_gender_test = data_test['Studentsex']
df_label_train = data_train['Dropout']
df_label_test = data_test['Dropout']
meals_train = np.array(df_meals_train)
meals_test = np.array(df_meals_test)
age_train = np.array(df_age_train)
age_test = np.array(df_age_test)
```

```
readbook_train = np.array(df_readbook_train)
readbook_test = np.array(df_readbook_test)
parent_discussion_train = np.array(df_parent_discussion_train)
parent_discussion_test = np.array(df_parent_discussion_test)
checkbook_train = np.array(df_parent_checkbook_train)
checkbook_test = np.array(df_parent_checkbook_test)
gender_train = np.array(df_gender_train)
gender_test = np.array(df_gender_test)
x_train = np.array([meals_train, age_train, readbook_train, parent_discussion_x_test =
np.array([meals_test, age_test, readbook_test,
parent_discussion_#===============Transposing the x_train and
x_test==============
x_train = x_train.T
x_test = x_test.T
X_train = pd.DataFrame(x_train)
X_test = pd.DataFrame(x_test)
X_train.head(3)
X_test.head(3)
model = ensembleclassifier()
model.fit(X_train, df_label_train)
predicted = model.predict(X_test)
accuracy = accuracy_score(df_label_test, predicted)
accuracy = accuracy*100
print("The accuracy of the model is {}%".format(accuracy))
from sklearn.externals import joblib
joblib.dump(model, "MODEL.pkl")
```